1991

Determining the Number of Principal Components: A Comparison of Two Methods

Cheryl A. Eaton
University of Rhode Island

DETERMINING THE NUMBER OF PRINCIPAL COMPONENTS:

A COMPARISON OF TWO METHODS

BY

CHERYL A. EATON

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

1991

ABSTRACT

The accuracy and variability of ten methods which determine the number of components to retain in a principal components analysis were examined. The methods consisted of three variations of the minimum average partial correlation method, six variations of parallel analysis, and the eigenvalue greater-than-one rule. The methods were investigated under different levels of five factors: sample size, component saturation, number of variables, number of variables per component, and the presence of unique items.

The eigenvalue-greater-than-one rule was the least accurate and most variable of all the methods. In every combination of the five factors, this method overestimated the number of components to retain. Both the parallel analysis method and the minimum average partial correlation method were found to be extremely accurate across a variety of combinations of the five factors. Alternate ways of implementing these two methods were found to be more accurate and less variable than the original version proposed for each method.

Component saturation and the number of variables per component were found to have the greatest effect upon the accuracy of all the methods. Higher saturation and more variables per component resulted in greater accuracy and less variability. Fewer variables also resulted in greater accuracy across all methods. The effect for sample size and unique items was not as notable or consistent across all methods.

ACKNOWLEDGEMENT

I would especially like to thank my major professor and advisor, Dr. Wayne Velicer, who first interested me in the topic of this study. His guidance, encouragement, and patience have been invaluable to me in completing this research.

I thank the members of my committee for their advice and thoughtful suggestions. I would also like to acknowledge the cooperation of the Academic Computer Center at the University of Rhode Island which provided the resources necessary to conduct this study, with special thanks to Charlene Yang.

Lastly, to my friends whose continued support and encouragement have been unwavering, I thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Principal components analysis is a procedure which is often employed to reduce a set of p observed variables into a smaller set of m derived variables or components ($m < p$). One area of continued investigation in this data reduction procedure is the determination of the optimal number of components to retain, i.e. the correct value for m.

Principal components analysis may be presented as an eigen decomposition into characteristic roots and vectors of the p x p sample correlation matrix,

[1] $\mathbf{R} = \mathbf{L}\ \mathbf{D}_\lambda\ \mathbf{L}'$

where **L** is an orthogonal matrix of columns of eigén vectors (weights) and $\mathbf{D}_\lambda$ contains the eigen roots (eigenvalues) representing the variance accounted for by each component. It then follows that

[2] $\mathbf{R} = \mathbf{L}\ \mathbf{D}_\lambda^{1/2}\ \mathbf{D}_\lambda^{1/2}\ \mathbf{L}'$

and

[3] $\mathbf{A} = \mathbf{L}\ \mathbf{D}_\lambda^{1/2}$

where **A** is the component pattern matrix. Thus, principal components analysis is equivalent to factoring the correlation matrix into a product of the pattern matrix and its transpose.

The main characteristic of principal components analysis is that each component is selected so that it accounts for the maximum possible variance of the variables. A second characteristic is that each component is extracted so it is uncorrelated, or orthogonal, to previously extracted

components.

The first principal component is defined as the linear combination of the p observed variables

[4] $Y_1 = b_{11}X_1 + b_{21}X_2 + \ldots + b_{p1}X_p$

under the constraint that the weight vector is of unit length

[5] $\underline{b}'_1 \underline{b}_1 = 1$

where the b_1 coefficients are elements in the eigen vector of weights associated with the greatest eigen root. Each of the succeeding principal components

[6] $Y_i = b_{1i}X_1 + b_{2i}X_2 + \ldots + b_{pi}X_p$

are defined under the constraints

[7] $\underline{b}'_i \underline{b}_i = 1$

and

[8] $\underline{b}'_i \underline{b}_j = 0$

as the linear combination where the $\underline{b}_i$ coefficients are elements in the eigen vector associated with the ith greatest eigen root.

There are potentially p components that can be derived. In seeking a parsimonious solution, the components with the smaller eigenvalues are dropped from the final solution since those components account for less variance. The problem then arises of how to determine which of the components should be retained and which components should be dropped. Although many different methods have been proposed, no one method has gained universal acceptance. The methods primarily differ in how the eigenvalues are examined to make a determination of the best

number of components to be retained. The rationale, procedure, and accuracy for methods of determining the number of components is presented.

Alternative Procedures: General

A commonly used method for determining the number of components to retain in a principal components analysis is the eigenvalue greater-than-one rule proposed by Kaiser (1960). With this method, all components with an eigenvalue greater than unity are retained. Guttman (1954) originally presented this rule as providing the lower bound of the number of common factors of a correlation matrix in the population. He did not suggest its use as a basis for determining the number of factors. Gorsuch (1983) criticized the applied use of the criterion for determining *the number* of factors rather than determining *the lower bound* for the number of factors. The acceptance of the rule as providing the lower bound was questioned by Schonemann (1990), who established that the logic that the rule provides the lower bound is not valid.

The use of the eigenvalue-greater-than-one rule has been supported by the intuitively appealing argument that one would only want to retain factors which account for more variance (greater than 1.0) than the original variable. Another rational for the use of this rule is the statement by Kaiser (1960) that components with eigenvalues less than 1.0 will have negative reliability. However, Cliff (1988) demonstrated

that this statement is false; the reliability of the components cannot be determined by the size of the eigenvalues. Cliff (1988) proposes that Kaiser's statement was based upon the incorrect application of the Kuder-Richardson 20 (K-R 20) formula for the reliability of a composite. In actuality, the reliability of a component is determined by the reliability of the measures.

The Kaiser criterion is the default for statistical software packages such as SPSS and BMDP. Although the eigenvalue-greater-than-one rule criterion is a very popular method, it has been shown to lead to overextraction of the number of components (Cattell & Jaspers, 1967: Hubbard & Allen, 1987: Lee & Comrey, 1979; Linn, 1968; Revelle & Rocklin, 1979; Yeomans & Golder, 1982; Zwick & Velicer, 1982, 1986). Typically, the number of components retained by this rule is related to the number of variables in ratios ranging from 1/3*p to 1/5*p rather than the actual structure of the data in situations with low communalities (Zwick & Velicer, 1982). Gorsuch (1983) suggests this method is appropriate as an approximate estimation of the number of factors in cases with less than 40 variables, a large N, and an expected number of factors between 1/3 and 1/5 of the number of variables. Hubbard and Allen (1987) state the routine use of this method is no longer justified, and Zwick and Velicer (1982, 1986) did not recommend using this procedure at all.

The scree test, another commonly used procedure for

determining the number of components to retain, was proposed by Cattell (1966). With this method, one plots the eigenvalues and examines the plot to find where a break occurs. At the point of the break, the number of components is indicated. The eigenvalues above the break indicate common components, while those below the break represent error variance. Problems arise with this method when there is not an obvious break or when there are several breaks, both of which lead to a more subjective judgement of the appropriate number of components. Many studies have found this method to be reasonably effective in suggesting the correct number of components to retain (Cattell & Jaspers, 1967; Cattell & Vogelmann, 1977; Cliff, 1970; Linn, 1968; Tucker, Koopman, & Linn, 1969; Zwick & Velicer, 1982). Zwick and Velicer (1982) found the scree test to be the most accurate of four methods for determining the number of components, especially in situations with large sample sizes and with strong components. Hakstian, Rogers, and Cattell (1982) found the scree test to be less accurate with low communality data, which resulted in an overidentification of the number of factors. The scree test has also been found to be less accurate with smaller sample sizes (Cliff & Hamburger, 1967; Linn, 1968). In a later study, Zwick and Velicer (1986) found the scree test to be less accurate than several other methods when more complex patterns such as those which included unique and complex items were considered. With complex patterns, which are more commonly seen in applied

situations, the scree test was found to be more variable and less accurate.

Another procedure for determining the number of components is Bartlett's (1950, 1951) test of significance of the residuals. The null hypothesis of the test is that, after the first m components are removed, the remaining eigenvalues are equal. In practice, one continues to remove components until the null hypothesis fails to be rejected. Horn and Engstrom (1979) discuss the similarities between Cattell's scree test and Bartlett's significance test. Both methods are based upon the same rationale, with an examination of the contribution of the remaining components after m components have been extracted. Horn and Engstrom (1979) express a preference for the more explicit method of the significance test over the subjective method of the scree test, although they state the scree test is useful. Not all agree with this preference. Gorsuch (1973) reports that Bartlett's significance test indicates the maximum, not necessarily the actual, number of components to extract and that it leads to the extraction of too many smaller, often trivial components. In a study comparing five different methods, Hubbard and Allen (1987) reported that Bartlett's test overestimated the number of components to retain. Zwick and Velicer (1982, 1986) found the accuracy of Bartlett's significance test decreased with smaller sample sizes, and was less accurate and more variable than the scree test.

Other methods have been proposed (Everett, 1983; Horn, 1965; Revelle & Rocklin, 1979; Velicer, 1976) to determine the number of factors or components to retain. Two of the most promising methods (Crawford & Koopman, 1973; Hubbard & Allen, 1987; Humphreys & Montanelli, 1975; Zwick & Velicer, 1982, 1986) are parallel analysis (Horn, 1965) and the minimum average partial (MAP) correlation method (Velicer, 1976). The next two sections will review these two procedures in detail.

Parallel Analysis (PA) Procedures

In 1965, Horn introduced the parallel analysis method for determining the number of factors. A set of random data correlation matrices, with the same number of variables and subjects as the observed data, is generated. The average of the eigenvalues across the set of random data matrices is calculated. The eigenvalues of the observed data are then compared to the averaged eigenvalues of the random data. Components are retained as long as the eigenvalue of the observed data exceeds the eigenvalue of the random data.

One problematic area is the determination of how many random data correlation matrices should be included. Although Horn (1965) used one random data correlation matrix in his introduction of parallel analysis, he proposed that the averaged eigenvalues should give the appropriate curve when the number of matrices is "reasonably large". Crawford and Koopman (1973) found no significant difference in the accuracy

of parallel analysis with eigenvalues from one random correlation matrix as compared to the averaged eigenvalues across 100 random correlation matrices. Longman, Cota, Holden, and Fekken (1991) found that, as expected, accuracy was greater with 40 than 3 replications. However they report that the difference in accuracy was not significant, and recommend parallel analysis based upon a few replications as a highly accurate procedure.

Due to the difficulty encountered when implementing this method, much recent work has been focused on developing alternatives to avoid the necessity of generating multiple correlation matrices. One alternative is the development of regression equations for predicting the random eigenvalues.

Montanelli and Humphreys (1976) introduced the first regression equation for use in determining the number of factors to retain in principal axes factor analysis. The equation predicts the common (base 10) logarithms of the latent roots of random correlation matrices, with squared multiple correlations (SMC) on the diagonal. The regression equation is given as:

$$[9] \qquad \log \lambda_i = a_i + b_{Ni} \log(N - 1) + b_{pi} \log\{(p(p-1)/2)-(I-1)p\}$$

where a is the intercept; b_{Ni} and b_{pi} are regression coefficients; N is the number of observations; p is the number of variables; and I is the ordinal position of the eigenvalue.

This equation estimates about half of the eigenvalues.

The first regression equation for use with principal components analysis was developed by Allen and Hubbard (1986). Their study varied the number of variables from 5 to 50 in steps of 5. The number of subjects included 30, 60, 90, 120, 240, 500, and 1000. All N and p combinations which satisfied the restriction that $N > 3p/2$ were examined. For cases with a sample size less than 240, 50 replications were employed. For those cases with a sample size of 240 or more, 30 replications were used. They presented the following equation for predicting the natural logarithms of latent roots of random data correlation matrices with unities on the diagonals:

$$[10] \quad \ln \lambda_i = a_i + b_i \ln(N-1) + c_i \ln\{(p-I-1)(p-I+2)/2\} + d_i \ln(\lambda_{i-1})$$

where a is the intercept; b, c, and d are the regression weights; N is the number of observations; p is the number of variables; I is the ordinal position from 1 to (p-2); and λ_0 equals 1. This equation predicts all eigenvalues except the last two. It is appropriate for situations with up to 50 variables.

Longman, Holden, and Fekken (1991) report the observance of anomalies with the Allen and Hubbard (1986) equation. They describe situations in which the eigenvalues predicted from the Allen and Hubbard (1986) regression equation continually

decrease as expected, until a point where the eigenvalues then increase. Longman, Cota, Holden, and Fekken (1989a) introduced a second regression equation for predicting eigenvalues from random data correlation matrices with principal components analysis. They suggest it is easier to calculate, and yields improved results as compared to the equation of Allen & Hubbard (1986). The values for the number of variables were 5, 10, 15, 20, 25, 35, and 50. The number of subjects included 50, 75, 100, 125, 150, 175, 200, 300, 400, and 500. After considering several variations of regression equations, the most accurate results were found with the following equation for predicting the natural logarithms:

$$[11] \quad \ln(\lambda_i) = a_i \ln(N_i) + b_i \ln(p_i) + c_i \{\ln(N_i) \ln(p_i)\} + d_i$$

where a, b, c, and d are the regression weights; N is the number of observations; p is the number of variables; and I is the ordinal position from 1 to (p-2). If p is 35 or less, the equation predicts p-2 eigenvalues. If p is greater than 35, then it is recommended that only the first 33 eigenvalues be predicted. Longman, Cota, Holden, and Fekken (1989a) compared this equation with the Allen and Hubbard equation on five combinations of N and p. For all five combinations, the Longman, Cota, Holden, and Fekken equation was more accurate than the Allen and Hubbard equation.

A third regression equation for predicting the

eigenvalues of random correlation matrices was proposed by Lautenschlager, Lance, and Flaherty (1989). The number of variables ranged from 5 to 50 in increments of 5. The number of subjects included 50, 75, 100, 150, 200, 300, 400, 500, 750, and 1000. Those N and p combinations which met the N > 3p/2 restriction were examined. The addition of a p/N term to the Allen and Hubbard equation improved the accuracy of the prediction of the first, and therefore subsequent, eigenvalues. This equation is:

$$[12] \quad \ln \lambda_i = a_i + b_i \ln(N-1) + c_i \ln\{(p-I-1)(p-I+2)/2\} + d_i \ln(\lambda_{i-1}) + e_i\, p/N$$

where a is the intercept; b, c, d, and e are regression weights; N is the number of observations; p is the number of variables; and I is the ordinal position from 1 to (p-2). This equation predicts up to 48 eigenvalues.

A second alternative to generating the multiple random data correlation matrices required to conduct parallel analysis is linear interpolation of tabled eigenvalues developed by Lautenschlager (1989a). These tables were generated by averaging the eigenvalues across random correlation matrices for 10 values of p, ranging from 5 to 50, and 12 values of N, ranging from 50 to 2000. The number of generated matrices was 100 when p > 10 and 200 for p <= 10. Lautenschlager (1989a) suggests this method is generally more

accurate than both the Allen and Hubbard (1986) regression equation and the Lautenschlager, Lance, and Flaherty (1989) regression equation.

Minimum Average Partial (MAP) Procedures

An alternative to the parallel analysis method of determining the number of components to retain is the minimum average partial correlation, or MAP, method (Velicer, 1976). The MAP procedure was developed for use with principal components analysis. As each component, m, is partialled out of the correlation matrix, the resulting partial correlation matrix is calculated. For each value of m, the average of the squared correlations of the partial correlation matrix is computed. The number of components to retain is indicated at the point where the average squared partial correlation reaches a minimum.

The matrix of partial correlations is obtained by first computing the partial covariance matrix,

$$[13] \quad \mathbf{C} = \mathbf{R} - \mathbf{A}\,\mathbf{A}'$$

where **C** is the partial covariance matrix, **R** is the correlation matrix, and **A** is the pattern matrix. The partial correlation matrix is then computed

$$[14] \quad \mathbf{R}^{*} = \mathbf{D}'\,\mathbf{C}\,\mathbf{D}$$

where $\mathbf{R}^*$ is the matrix of partial correlations and $\mathbf{D}$ is the diagonal of the partial covariance matrix. The equation for calculating the MAP criterion from the matrix of partial correlations is given as

$$[15] \quad MAP_m = \Sigma \Sigma (r_{ij}^*)^2 / (p(p - 1))$$

where r_{ij}^* is the value in row i and column j of the matrix of partial correlations and p is the total number of observed variables.

Velicer (1976) established that as common variance is partialled out of the matrix for each successive value of m, the MAP_m criterion will continue to decrease. At the point where the common variance has been removed and only unique variance remains, the MAP_m criterion will begin to rise. If one examines the general formula presented by Velicer (1976) for computing a partial correlation,

$$[16] \quad r_{ij.y} = \frac{r_{ij} - (r_{iy} * r_{jy})}{(1-r_{iy}^2)\ (1-r_{jy}^2))^{1/2}}$$

one sees that the partial correlation coefficient will decrease as long as the numerator decreases more than the denominator. When the denominator begins to decrease more than the numerator, the partial correlation coefficient will then begin to increase. The latter would occur when a high correlation of a component with one variable and low

correlations with all other variables was present; this situation characterizes the extraction of a unique variable.

The MAP method was found to be comparable to the scree test and more accurate (Zwick & Velicer, 1982) than the eigenvalue greater-than-one rule and Bartlett's significance test for the number of components. In a later study with more complex data, Zwick and Velicer (1986) demonstrated that both the MAP criterion and parallel analysis were more accurate than the eigenvalue greater-than-one rule, Cattell's scree test, and Bartlett's significance test for the number of components.

An alternative way of computing the MAP criterion is by computing the trace, or sum, of the eigenvalues of the squared matrix of partial correlations as each of the m components are extracted. The number of variables is subtracted from the trace, and the remainder is then divided by the number of off-diagonal elements in the matrix, (p(p-1)). The resulting value equals the MAP criterion. Thus,

$$[17] \quad MAP_m = \Sigma \Sigma (r_{ij}^*)^2 / (p(p-1)) = \text{Trace } \mathbf{R}^{*2} - p / (p(p-1))$$

Other variations of this procedure which might display a sharper point for the directional change from the removal of common to unique variance have not yet been explored. Two aspects of the current MAP criterion are the use of the trace

of a matrix and the squared power of the matrix.

The trace of a matrix is one of three summary statistics (the trace, the determinant, and the largest root) which are used to summarize matrices in multivariate procedures. The largest root is the largest eigenvalue. The determinant is the product of the eigenvalues. For each value of m, the matrix of partial correlations will contain m zero eigenvalues, which will result in a value of zero for the determinant. Therefore, the determinant is excluded from consideration as a variation of the MAP criterion. The two remaining matrix summary statistics, the trace and the largest root, are both viable options.

The power of the matrix under consideration may also be varied. The current calculation of the MAP criterion is based on the squared matrix. In a search for variations which may give a sharper directional change at the minimum value, the first, second, and fourth powers all hold potential as possible alternatives of calculating the criterion.

Purpose of the Study

Two promising methods of determining the number of components, parallel analysis and the minimum average partial correlation, have been presented. This study was conducted to evaluate the performance of several variations of each of these methods. The methods were evaluated across multiple factors known to affect the performance of the methods.

Method

This study examined the accuracy of six alternate methods of parallel analysis and three variations of the MAP procedure in determining the number of components to retain in a principal component analysis. Several variations, as well as the original method, of each procedure were included. The eigenvalue-greater-than-one rule was also included.

Decision Rules Evaluated

The following methods of parallel analysis were examined: 1) R5 - the generation of 5 random correlation matrices and averaging of eigenvalues across the matrices, 2) R100 - the generation of 100 random correlation matrices and averaging of eigenvalues across the matrices, 3) AH - the regression equation presented by Allen and Hubbard (1986), 4) LCHF - the regression equation proposed by Longman, Cota, Holden and Fekken (1989), 5) LLF - the regression equation introduced by Lautenschlager, Lance, and Flaherty (1989), and 6) TAB - interpolation from the tables of mean criterion eigenvalues provided by Lautenschlager (1989a).

Three versions of the MAP procedure were included. The versions differ in the power of the matrix under consideration and in the matrix summary statistic. The three versions are: 1) TR2 - the previously studied version consisting of the trace of the matrix of squared partial correlations, 2) TR4 - the trace of the matrix of partial correlations to the fourth

power, and 3) LR1 - the largest root of the matrix of partial correlations. The trace of the matrix of partial correlations is omitted since the value would always equal p, the number of variables. Only one of the possible three versions of the largest root of the matrix of partial correlations was employed. Since the first, second, and fourth power of the largest root mathematically produce the same result, the second and fourth power are redundant and were omitted.

One additional method included in this study was the eigenvalue greater-than-one rule. As the most commonly employed method in applied use, this method served as a baseline to evaluate the alternative methods.

Design

Several factors are known to influence the accuracy of these methods in determining the number of components to retain in a principal components analysis. These factors include sample size (N), the number of variables (p), the ratio of the number of variables per component (p:m), and the component saturation (CS). These factors were varied in the current study in an attempt to determine the performance of the methods under selected conditions.

Values were selected for the levels of these factors according to two criteria: 1) those which best represent values which are found in applied research settings and 2) those which have been demonstrated to differentiate the

accuracy of decision methods in other Monte Carlo studies. Three levels of the number of variables, the sample size, and of component saturation were included. Two values for the ratio of the number of variables per component were selected. The next section presents the basis for the selection of these values.

Selection of the Levels of Factors Influencing Method Accuracy

The number of variables, p, was set at 24, 48, and 72. Since principal components analysis is a data reduction procedure, it is not typically employed on data with very few variables. The value of 24 was selected to reflect a small data set, 48 a moderate data set, and 72 a larger data set. These values fall within the range of values which have been investigated extensively (Anderson, Acito, & Lee, 1982; Hakstian, Rogers, & Cattell, 1982; Lee & Comrey, 1979; Velicer & Fava, 1987; Velicer, Peacock, & Jackson, 1982; Zwick & Velicer, 1986).

Anderson, Acito, and Lee (1982) reported that sample size was one of two factors which had the most effect on the accuracy of the decision methods they examined. Three levels of sample size were selected to be included in this study: 75, 150, and 300. The value of 75 is considered to be a low sample size, while 150 has been recommended (Velicer, Peacock, & Jackson, 1982; Guadagnoli & Velicer, 1988) as a minimum to provide adequate results when the component saturation and the

number of variables per component are sufficient. When the number of variables per component or the component saturation are at minimum ranges, the larger sample size of 300 is recommended (Guadagnoli & Velicer, 1988).

The levels of the number of variables, p, and the number of components, m, were selected to ensure that the variables per component ratio was held constant at 4:1 and 8:1. The ratio of 4:1 is viewed as just over the minimum number of variables needed to define a component, and the ratio of 8:1 represents a moderately strong component. This p:m ratio has repeatedly (Guadagnoli & Velicer, 1988; Velicer & Fava, 1987; Yeomans & Golder, 1982; Zwick & Velicer, 1982, 1986) been found to influence the accuracy of the results, with more variables per component producing more stable values. For p = 24, m was set at 3 and 6; for p = 48, m was selected to be 6 and 12; and for p = 72, m was set at 9 and 18.

The magnitude of the component loadings, or saturation, has repeatedly been found to be one of the factors having the greatest effect on accuracy within principal components analysis. Guadagnoli and Velicer (1988) found that component saturation was one of the two factors which most influenced the stability of the component solution. Hakstian, Rogers, and Cattell (1982) and Anderson, Acito, and Lee (1982) both reported that the decision methods they examined all performed best with high component loadings. Yeomans and Golder (1982) found the performance of the Kaiser criterion differed under

varied levels of saturation. Velicer and Fava (1987) reported that component saturation, as well as the p:m ratio, had a large effect on pattern reproduction, with high saturation and a high ratio providing excellent pattern reproduction. Zwick and Velicer (1982, 1986) found that increased component loadings of .80 as compared with .50, had an impact on the accuracy of the decision methods. The three levels of component saturation included in this study were selected to provide for an assessment over a range of values. The levels selected were .40, .60, and .80 which represent low, medium, and high component loadings, respectively.

An additional factor which may affect the accuracy of these decision methods is the presence of items which are unique. Unique items have only one nonzero component loading and no other items load on that component. The presence or absence of unique items (U) will be an additional factor varied in this study. The values of the resulting 3 X 3 X 3 X 2 X 2 design are displayed in Table 1.

Data Generation

There are two major approaches to examining the accuracy of decision methods for the number of components. The first approach is to apply a method to either newly observed data sets or well established, "classic" data sets. This method was employed by Cattell (1966), Crawford (1975), Horn (1965), Humphreys and Montanelli (1975), Lee and Comrey (1979), and

Velicer (1976).

The second approach is to simulate data sets with a specified number of components (Anderson, Acito, & Lee, 1982; Crawford & Koopman, 1973; Everett, 1983; Hakstian, Rogers, & Cattell, 1982; Revelle & Rocklin, 1979; Tucker, Koopman, & Linn, 1969; Yeomans & Golder, 1982; Zwick & Velicer, 1982,1986). The value of the number of components, m, is both known and under the control of the experimenter. This approach was selected for this study. It allowed for an evaluation of the decision methods under a variety of levels of the number of subjects, the number of variables, the number of components, and the component saturation. Population correlation matrices were generated with the number of components as specified above (m = 3 and 6 for p = 24; m = 6 and 12 for p = 48; and m = 9 and 18 for p = 72). Each of the decision methods were then evaluated on whether the correct number of components was indicated for the matrix being analyzed.

The procedure previously employed by Guadagnoli and Velicer (1988,1991), Velicer, Peacock, and Jackson (1982), Velicer and Fava (1987), and Zwick and Velicer (1982, 1986) was utilized. The population correlation matrices were generated in the following manner: 1) the component pattern, **A**, based upon the combination of values for p, m, and CS was created, 2) the pattern matrix was multiplied by its transpose (**A**′) which resulted in a matrix $\mathbf{R}_1$ ($\mathbf{R}_1 = \mathbf{A}\,\mathbf{A}'$), and 3) values

of 1.0 were substituted in the diagonal of the $\mathbf{R}_1$ matrix which added error and created a correlation matrix, $\mathbf{R} = \mathbf{R}_1 + \mathbf{D}^2$, of full rank. Table 2 illustrates this sequence of generating a population matrix with an example of 6 variables, with two components and moderate (.6) component saturation. The pattern matrix, **A**, the resulting $\mathbf{R}_1$ matrix, the computed $\mathbf{D}_2$ matrix, and the resulting population correlation matrix, **R** are displayed.

Procedure

Population correlation matrices were generated for each of the 108 combinations of the 3 X 3 X 3 X 2 X 2 design. For each of the population correlation matrices, five sample correlation matrices were then generated employing a program by Montanelli (1975). This is based on a method proposed by Odell and Feiveson (1966). The number of samples to generate was evaluated by Guadagnoli and Velicer (1991) and five was considered adequate.

For those matrices which were generated with unique items added, the total number of items was increased to maintain the 8:1 and 4:1 variables per component ratios. The increased number of variables was 27 and 30 for p = 24, 54 and 60 for p = 48, and 81 and 90 for p = 72. For the 30 matrices where N = 75, the number of variables then exceeded the number of subjects, and these conditions were excluded from analysis.

A principal components analysis was performed on each of the resulting 510 correlation matrices. The number of

components retained by each of methods was then computed. The following section presents the procedures used to compute the number of components for each method.

Computation of M

The eigenvalues of the correlation matrix were examined to determine the number of components to retain. The value for the number of components retained by the eigenvalue greater-than-one rule was calculated by counting the number of eigenvalues that were greater than 1.0.

The number of components retained by the three MAP variations was computed by adapting the CAX (Component Analysis Extended) Fortran program (Velicer, Fava, Zwick, & Harrop, 1990). The program was expanded to include the calculation of the trace of the matrix of partial correlations to the fourth power and the largest root of the matrix of partial correlations.

For the six parallel analysis methods, the eigenvalues of the simulated data were compared to the eigenvalues produced by the random data correlation matrices, the three regression equations, and the interpolated tabled values provided by Lautenschlager (1989a). Several computer programs were utilized to determine the number of components for these methods. For random data parallel analysis, the generation of the random correlation matrices and the averaging of the random data eigenvalues was performed using the PAM Fortran

program presented by Longman, Cota, Holden, and Fekken (1989b). The implementation of the regression equations for estimating the eigenvalues of the equations presented by Allen and Hubbard and by Longman et al. were completed with the PAR Fortran program (Holden, Longman, Cota, & Fekken, 1989). This program was adapted to also estimate the eigenvalues for the equation proposed by Lautenschlager, Lance, and Flaherty (1989).

Results

Number of Correlation Matrices Examined

As stated previously, 30 of the possible 540 matrices where $N < p$ were eliminated from consideration. The resulting number of matrices to be examined was 510. The eigenvalue-greater-than-one rule, the three minimum average partial correlation methods, and the two random data parallel analysis methods were examined on the 510 matrices. Due to limitations of the methods, the other parallel analysis methods were examined on less than the 510 matrices.

Lautenschlager's tabled eigenvalues method was analyzed on 76% of the matrices. Of the 510 matrices, 390 were able to be examined. Tables were not provided for $N = 75$ with $p = 54$ and 60, or for $N = 75$, 150, and 300 with $p = 81$ and 90. These values of p occurred when unique items were added to the data.

The three parallel analysis regression equation methods

were examined on 53% of the matrices. These methods were examined on 270 of the possible 510 matrices. The application of the equations is limited to the values of N and p on which the equations were developed. Since 50 is the maximum value of the number of variables for the three regression equations, the conditions of p = 72 with no unique items and N = 75, 150, and 300 had to omitted from examination. In addition, conditions were excluded for N = 75, 150, and 300 with p = 54, 60, 81, and 90 when unique items were added.

Measures of Method Performance

Deviation scores were computed to evaluate each of the ten decision methods. The accuracy of the number of components given by each method was computed by subtracting the number of components indicated (m estimated) for the simulated data from the correct number of components (m) for each sample of that combination of N, p, p:m, CS, and U. A value of 0 indicated the method was correct for that case. The average of the deviation scores was then computed. Values of zero indicate that the method was accurate on the average. Negative values indicate that the method overestimated, on the average, while positive values indicate that the method underestimated the correct number of components on the average.

The standard deviation of the averaged deviation scores was also calculated as a measure of variability for each method. A high standard deviation indicates considerable

variability in estimating m in the different conditions.

A third measure was also computed to evaluate the performance of the methods. The percent of the total cases available where the estimate of the number of components was exactly correct was computed.

Overall Performance of the Decision Methods

These three measures of the performance of the decision methods were first examined with the data collapsed across all levels of the five factors. These data are displayed in Table 3. Both the parallel analysis methods and the minimum average partial correlation methods slightly underestimated the correct number of components but were generally quite accurate. An exception was the Allen and Hubbard (1986) regression equation which slightly overestimated the value for m. The eigenvalue-greater-than-one rule consistently overestimated the number of components.

Within the parallel analysis methods, the three regression equations were generally the most accurate, the least variable, and had the highest percent of correct estimations of m. The averaged deviation scores for the three equations ranged from -1.16 to .60. The use of the tabled eigenvalues was next in performance with an average deviation score of 1.50, followed by the random generation method (mean = 2.11, 2.13 for 100 and 5 replications respectively). One exception to this pattern of results can be noted. The Allen

and Hubbard (1986) equation was the most variable of the parallel analysis methods (S.D. = 8.87).

The performance of the three minimum average partial correlation methods is also presented in Table 3. The three MAP versions displayed similar patterns for both accuracy and variability. In general, the largest root of the matrix of partial correlations was the most accurate and least variable of the three versions with an average deviation score of .59. The trace of the matrix of partial correlations to the fourth power (mean = 1.89) was next in overall performance. The trace of the squared matrix of partial correlations was the least accurate and most variable, (mean = 2.46). A different pattern of results was observed for the three MAP versions on the performance measure of percent of correct estimations. The trace of the matrix of partial correlations to the fourth power had the highest percent correct, 72.5%, followed by the trace of the squared matrix of partial correlations with a percent correct value of 66.9%. The largest root version displayed a considerably lower percent of correct estimations, 45.9%. Figures 1, 2 and 3 portray the patterns of the MAP criterions for each of the three MAP variations as the first 15 components are partialled out of the correlation matrix.

The eigenvalue-greater-than-one rule was the least accurate of all the methods. The value for m was markedly overestimated by this rule with an average deviation score of -6.27. The percent of estimations of m which were correct was

substantially lower than all of the other methods (22.0%).

The accuracy and variability of the methods was also examined separately for each of the five factors (N, p, p:m, CS, U) varied in this study. The effect of each factor was examined, after collapsing the data across the other four factors. The next three sections present these results separately for the six parallel analysis methods, for the three minimum average partial correlation methods, and for the eigenvalue-greater-than-one rule.

Parallel Analysis Methods

The effect of each of the five factors varied in this study on the performance of the six parallel analysis methods was examined. Table 4 presents the effect of varying the levels of component saturation on the accuracy, variability, and percent of correct estimations of m. For all six parallel analysis methods, the accuracy increased, the percent of correct estimations of m increased, and the variability decreased as saturation was increased. At low (.40) saturation, the three regression equation methods were the most accurate and had the greatest number of correct estimations of the number of components, with an average deviation score of -1.54 to 2.83. Although high in accuracy, the Allen and Hubbard (1986) equation was also notably high in variability. The tabled values were next in accuracy (mean = 4.25) and percent of correct estimations, and the random

generation method was the least accurate of the methods at low saturation. At moderate (.60) saturation, the Allen and Hubbard (1986) equation was considerably less accurate and more variable than the other methods. The other five parallel analysis methods displayed a substantial increase in accuracy, increase in the percent of correct estimations of m, and decrease in variability, especially the Longman, Cota, Holden, and Fekken (1989) and Lautenschlager, Lance, and Flaherty (1989) regression equations and the tabled values. When the component saturation was high (.80), all the parallel analysis methods were extremely accurate, with low variability. The average deviation scores ranged from .00 to .14.

The effect of the ratio of the number of variables per component can be seen in Table 5. Excluding the Allen and Hubbard (1986) regression equation, the methods exhibited considerably greater accuracy, lower variability, and a higher percent of correct estimations with the larger ratio of variables per component ratio, (mean = .10 to .40). The Longman, Cota, Holden, and Fekken (1989) equation, the Lautenschlager, Lance, and Flaherty (1989) equation, and the tabled values performed best, followed by the random data generation method. This pattern of results was consistent across both levels of the p:m ratio and all three measures of method performance. The Allen and Hubbard (1986) regression equation displayed a very different performance in accuracy and variability than the other methods. The accuracy was

better at the low p:m ratio and worse at the higher p:m ratio. The variability was substantially greater at both levels of the p:m ratio than the other methods. The percent of correct estimations of m was the same or slightly better as the other two regression equations at both levels of the p:m ratio.

Table 6 displays the accuracy, variability, and percent of correct estimations for the methods across the three levels of the number of variables. All methods were the most accurate (mean = -.12 to .66), exhibited the greatest percent of correct estimations, and least variability with the fewest number of variables, 24. At p = 24, the Longman, Cota, Holden, and Fekken (1989) regression equation was more accurate (mean = -.12) but also more variable than the other methods. At the moderate level of the number of variables, the Allen and Hubbard (1986) equation was less accurate (mean = -4.72) and considerably more variable than the other methods. There were no other notable differences in accuracy, variability, or the percent of correct estimations for the six methods. One confounding factor in examining the performance of the methods for the differing levels of the number of variables is the differing number of components at each level of p. The study was designed to maintain equal p:m ratios across the levels of p. To accomplish this, the number of components was allowed to vary across the levels of p (m = 3 and 6 for p = 24; m = 6 and 12 for p = 48; and m = 9 and 18 for p = 72). The superior performance of these methods at the lowest value of p

corresponds to the lowest number of components. As the number of components in the data was increased, the performance decreased.

The accuracy, variability, and percent of correct estimations for the methods under the three levels of sample size are displayed in Table 7. The accuracy and percent of correct values for m increased as sample size was increased from 75 to 150 to 300. At the highest sample size of 300, the three regression equations were extremely accurate, with average deviation scores ranging from -.20 to .71. The greatest variability for all methods was observed at the moderate level of sample size, N = 150. Generally, the regression equations displayed the best performance, followed by the tabled values, and lastly by the random data method. There were two exceptions to this pattern of results. The Allen and Hubbard (1986) equation was considerably less accurate (mean= -5.47) and more variable than the other methods at the lowest sample size, 75. At the moderate level of sample size, the Lautenschlager, Lance and Flaherty (1989) equation was notably more accurate (mean = .06) but also more variable than the other methods.

Table 8 presents the effect of the presence of unique variables on the accuracy, variability, and percent of correct estimations. The methods were more accurate, less variable, and had a higher percent of correct estimations when unique items were present, although for the random data generation

methods these differences were slight. Of the six parallel analysis methods, the regression equations were more accurate than the tabled values method, which was more accurate than the random data method. The Allen and Hubbard (1986) equation was less accurate with an average deviation score of -2.07 and substantially more variable than the other methods when unique items were not present in the data.

Minimum Average Partial Correlation Methods

The effect of the five factors varied in this study on the three minimum average partial correlation methods were examined. The accuracy, percent of correct estimations, and variability are presented for the five factors in Tables 4 through 8. As noted with the parallel analysis methods, the three factors which had the greatest effect on the accuracy of these three methods were component saturation, the variables per component ratio, and the number of variables. The methods displayed differing results for the three measures of performance. The largest root method was usually the most accurate and least variable of the three methods, but it also displayed the lowest percent of correct estimations of m. The performance of these three methods is presented next for each of the factors varied in the study.

The accuracy, variability, and percent of correct estimations for the three minimum average partial correlation methods under the varying levels of component saturation are

presented in Table 4. The accuracy and the percent of correct responses increased and variability decreased as saturation was increased. This effect for component saturation is the same as was observed with the parallel analysis methods. At low (.40) saturation, the largest root of the matrix of partial correlations was the most accurate and least variable, with an average deviation score of mean = 3.49. At moderate (.60) saturation, the trace of the matrix of partial correlations to the fourth power was the most accurate (mean = .09) and the least variable. When the component saturation was high, .80, all three methods were extremely accurate with low variability. The squared matrix of partial correlations demonstrated the greatest accuracy and least variability (mean = -.01) of the three methods at high component saturation, although the improvement over the trace of the matrix of partial correlations to the fourth power was negligible. The patterns for the percent of the estimations of m which were exactly correct differed from the accuracy and variability results. At the low and moderate levels of component saturation, the trace of the matrix of partial correlations to the fourth power displayed the highest percent of correct estimations. At both moderate and high saturation, the largest root version displayed a noticeably lower percent of correct estimations of m than the two trace MAP versions.

The effect of the number of variables per component ratio can be seen in Table 5. As seen with the parallel analysis

methods, all three methods were most accurate for the higher p:m ratio, (mean = -.88 to .52). At the higher p:m ratio, the trace of the matrix of partial correlations to the fourth power was the most accurate (mean = .08), least variable, and displayed the highest percent of correct estimations of the three methods. The trace of the squared matrix of partial correlations was next in performance at this p:m level. When examined for the lower variables per component ratio, the largest root method was the most accurate (mean = 2.06) and least variable, followed by the trace of the matrix of partial correlations to the fourth power. At both levels of the p:m ratio, the trace of the matrix of partial correlations to the fourth power had the greatest percent of correct observations of m, while the largest root had the lowest percent.

Table 6 displays the accuracy, variability, and percent of correct estimations of the methods across the three levels of the number of variables. The methods performed best when fewer variables were present. As stated in the section describing the parallel analysis results for the levels of p, the better performance with less variables coincides with less components to be identified. The largest root method was the most accurate and least variable at all three levels of the number of variables. The trace of the matrix of partial correlations to the fourth power resulted in the highest percent of correct estimations of m at each of the three levels of the number of variables. At 48 and 72 variables, the

largest root method had considerably fewer correct responses than both of the trace versions.

All three methods displayed different patterns of results at the varying levels of sample size (See Table 7). As sample size increased, the accuracy of the trace of the squared matrix of partial correlation method increased, while the accuracy of the largest root method decreased. For the trace of the matrix of partial correlations to the fourth power, accuracy increased slightly as sample size increased from 75 to 150, then decreased slightly as sample size increased from 150 to 300. At all three levels of sample size, the largest root version was considerably more accurate and less variable than the trace versions, but also had considerably less correct estimations of m.

Table 8 presents the effect of the inclusion of unique variables on the accuracy, variability, and percent of correct estimations. As observed with the parallel analysis methods, the methods generally performed better when unique items were added to the data. When unique items were both present and absent, the largest root method was the most accurate (mean = .62) and least variable of the three methods, while the trace of the matrix of partial correlations to the fourth power had the highest percent of correct estimations of m.

Eigenvalue-Greater-Than One Rule

In all cases, the eigenvalue-greater-than-one rule

overestimated the number of components to retain. The different levels of component saturation, the number of variables, and the inclusion of unique items had the greatest effect on the number of the overestimations by this method. As the component saturation was increased, the accuracy increased, the variability decreased, and the percent of correct estimations of m increased. The eigenvalue-greater-than-one rule was more accurate with a greater percent of correct estimations and less variation as the number of variables was reduced. The addition of unique items to the data resulted in the retention of twice as many components and a decrease in the percent of correct estimations from 41.5% to 0.0%. A slight increase in accuracy was noted at the highest level of sample size. In contrast to the other methods, the eigenvalue-greater-than-one rule was slightly more accurate and less variable with fewer variables per component.

The eigenvalue-greater-than-one rule gave the average number of components as approximately one third the number of variables (m=7.4 at p=24, m=15.8 at p=48, and m=23.3 at p=72). The association between m and p is presented in Table 9 collapsed across the five factors, and separately for the levels of component saturation and the p:m ratio. The overestimations of the eigenvalue-greater-than-one rule were greatest at low component saturation and the lower p:m ratio.

Patterns of Over and Under-Estimations For All Methods

Table 10 presents the percent of estimations of m, collapsed across all the conditions of the study, that were an overestimation or an underestimation of the correct value. For the parallel analysis and MAP methods, 12.6% to 32.4% of the estimations of m were underestimations. In addition, the Allen and Hubbard (1986) regression equation, the Lautenschlager, Lance, and Flaherty (1989) regression equation, and the two trace versions of the MAP procedure also occasionally overestimated the number of components. The largest root MAP version exhibited a different pattern. This method overestimated (36.1%) the value of m twice as often as it underestimated (18.0%). When the eigenvalue-greater-than-one rule was incorrect, it was always by overestimation (78%).

The percent of overestimations and underestimations are presented in Tables 11 through 13 for the three factors of the study which had the greatest effect upon the accuracy of the methods. These factors were component saturation (Table 11), ratio of the number of variables per component (Table 12), and the number of variables (Table 13).

For each of these three factors, the patterns of overestimations and underestimations for each of the parallel analysis methods was the same as when collapsed across all factors. When the estimation of m was incorrect, for all six methods it was generally by underestimation. The Allen and Hubbard (1986) and the Lautenschlager, Lance, and Flaherty (1989) regression equations also occasionally overestimated

the number of components.

The three MAP versions differed in the direction of the incorrect estimations of m. The trace versions primarily underestimated at low and moderate component saturation, at both levels of the p:m ratio, and at all three levels of the number of variables. At high saturation, when these methods were incorrect, it was by overestimation (1.8% and 8.2%). The largest root method exhibited a different pattern of overestimations and underestimations. The method showed a substantial percent of both overestimations and underestimations at low component saturation, the low p:m ratio, and at the lowest level of the number of variables. The largest root version primarily overestimated at moderate and high saturation and the higher p:m ratio of 8 variables per component. At both moderate and high levels of the number of variables, the method both overestimated and underestimated the value of m. As the number of variables increased to 48 and 72, the method overestimated considerably more often than it underestimated the correct number of components.

Comparisons of Observed and Predicted Eigenvalues

The regression equations indicate the number of components at the point where the value of the observed eigenvalue first falls below the corresponding value of the estimated eigenvalue, as depicted in Figure 4. Several other patterns of the observed data as compared to the estimated

data were observed.

One alternate pattern of the comparison of the observed and the estimated data is graphed in Figure 5. In this case, the higher value repeatedly alternated between the observed and the estimated data. The value for m was given at the first occurrence where the eigenvalue of the observed data was less than the corresponding eigenvalue of the random data.

Figure 6 portrays a second alternate pattern in which the observed eigenvalues remained greater than the estimated eigenvalues until all p-2 components were examined. In these cases, the number of components retained was p-2, the maximum number possible. There were 15 occurrences of this situation with the Allen & Hubbard equation and 5 occurrences with the Lautenschlager, Lance, and Flaherty equation. Table 14 lists the specific conditions of these 20 situations.

Since these cases of solutions of m = p - 2 can be viewed as inappropriate solutions which greatly affected the computed mean accuracy, the performance of the three regression equations was examined with solutions of m = p - 2 omitted from the accuracy calculations. Table 15 presents the deviation score means, standard deviations, and percent of correct estimations of m for the three regression equation methods, with the data collapsed across the five factors. The three regression equations were all extremely accurate, with average deviation scores of .89 to 1.03. The accuracy, variability, and percent of correct estimations are displayed

in Table 16 for the three factors which had the greatest effect on the performance of the methods: component saturation, the number of variables, and the p:m ratio. In all cases, the three equations display comparable accuracy and variability. The differences that existed when all cases were included in the computations are eliminated when the cases with estimations of m = p-2 are omitted. The decreased accuracy and increased variability of the Allen and Hubbard (1986) and the Lautenschlager, Lance, and Flaherty (1989) equations with all solutions included as compared to the Longman, Cota, Holden, and Fekken (1989) equation were no longer present.

Ambiguous Solutions for M

Zero is the minimum possible value for the number of components for all the parallel analysis and MAP methods. For the six parallel analysis methods, whenever the first randomly generated, estimated, or tabled eigenvalue is greater than the first eigenvalue of the observed data, the method indicates 0 components. Figure 7 depicts an example of an estimation of m = 0. The three minimum average partial correlation methods may also give 0 as a value for the number of components. The average correlation of the correlation matrix before any components are partialled out is used as the initial minimum value that is compared to the subsequent values of the partial correlation matrix. A value of 0 is given for m when the

partialling out of components does not result in a reduced value for the correlation matrix.

The situations in which the value of 0 occurred were consistent across all nine methods. The situations were all with lower component saturation and, except for three cases, with the low variables per component ratio. Each of the methods gave 0 for the value of m across all levels of sample size, number of variables, and presence of unique items.

The presence of estimations of 0 greatly affected the accuracy and variability computations. The number and percent of cases in which each method retained 0 components are given in Table 17. These cases occurred in 4.3% to 15.1% of all the cases examined. The largest root version of the MAP procedure had the fewest (4.3%) occurrences of 0. The trace of the matrix of partial correlations to the fourth power had considerably fewer 0 estimations (9.8%) than the trace of the squared matrix of partial correlations (15.1%). Of the parallel analysis methods, the regression equations had the fewest 0 cases (7.8% to 8.9%), followed by the tabled values method (9.5%). The random data generation method exhibited the most 0 cases of all the parallel analysis methods (11.2%).

Solutions of m = 0 are difficult to interpret. Users of these methods have no way to determine if the true solution is 0 components or if a problem existed (low saturation, low p:m ratio) and the method was unable to reach a solution. The data were therefore reanalyzed with solutions of m = 0 omitted from

the calculations of the three measures of performance. Since the solutions of m = p - 2 by the Allen & Hubbard (1986) and the Lautenschlager, Lance, & Flaherty (1989) regression equations are also considered invalid, these additional cases were omitted from the calculations. Table 18 presents the performance of the parallel analysis and MAP methods with these cases excluded. All methods were extremely accurate, with average deviation scores ranging from .21 to .99. The variability was consistently low across all the methods. The percent of correct estimations was high, (78.8% to 92.8%), with the exception of the largest root MAP method (48.0%).

Discussion

This study examined the performance of ten methods for determining the number of components to retain in a principal components analysis. The methods consisted of the eigenvalue-greater-than-one rule, six variations of Horn's (1965) method of parallel analysis, and three modifications of Velicer's (1976) minimum average partial (MAP) correlation method.

The methods were examined on simulated data which were created to reflect varied levels of sample size (3), number of variables (3), component saturation (3), number of variables per component ratio (2), and presence of unique items (2). For each of the resulting 108 conditions, five sample correlation matrices were created and a principal components analysis was performed.

There are four major considerations when evaluating methods for determining the number of components to retain. The rationale of the method, the ease of implementation, the accuracy of the method, and the number of cases under which the method is applicable are all key factors to examine when selecting a method. A preferred method should have a strong rationale, be readily available for implementation in applied settings, demonstrate a superior overall performance, and be applicable across a wide variety of conditions. The next three sections discuss the utility of the three methods evaluated in this study with respect to each of these considerations.

The Eigenvalue-Greater-Than-One Rule

I. Rationale

The rationale of the eigenvalue-greater-than-one rule has been challenged in recent years. As stated previously, the method was proposed as providing the lower bound of the number of components rather than the actual number of components. However, Schonemann (1990) has established that the logic of the rule providing the lower bound is not valid. A second rationale for the use of this method was that the eigenvalue must be greater than 1.0 for the reliability to be positive. Cliff (1988) demonstrated that the reliability of components is not determined from the size of the eigenvalues. These arguments considerably weaken the rationale of this method.

II. Implementation

The eigenvalue-greater-than-one rule is the easiest method for researchers in applied settings to implement. No specialized technical or statistical knowledge is required to use the method. The eigenvalue-greater-than-one rule is available on all major statistical software packages, and is usually the default method. This ease of implementation has contributed to the continued use of this method, even after numerous studies have demonstrated the poor performance of the method as compared to many other available methods.

III. Performance

The performance of the eigenvalue-greater-than-one rule was the poorest of all the decision methods evaluated in this study. It had the lowest accuracy and the lowest percent of correct estimations of all the methods. The eigenvalue-greater-than-one rule consistently overestimated the number of components to retain. There is no justification for the continued use of this method.

The overestimation of the number of components was consistent with other studies (Cattell & Jaspers, 1967: Lee & Comrey, 1979; Linn, 1968; Revelle & Rocklin, 1979; Yeomans & Golder, 1982; Zwick & Velicer, 1982, 1986). The number of components suggested by this rule was found to be related to the number of variables, 1/3 * p. These findings are consistent with Gorsuch (1983) and the Zwick & Velicer (1982)

study where the eigenvalue-greater-than-one rule was found to give the number of components to retain as 1/3 to 1/5 * p.

IV. Number of Cases

The eigenvalue-greater-than-one rule was applicable for all the conditions of this study. A strength of this method is that it may be applied to data based upon any value of N and p, the number of subjects and variables.

The Three MAP Methods

I. Rationale

The minimum average partial correlation procedures for determining the number of components to retain have a strong rationale. The MAP procedures identify the number of components at the point where the minimum value of the MAP criterion is observed. This occurs when all the common variance has been removed from the correlation matrix, and only unique variance remains. If one continues to remove components after the minimum is reached, the criterion then increases, indicating that unique variance is being removed from the correlation matrix. Thus, the MAP criterions separate common and unique variance, and retain only components consisting of common variance.

II. Implementation

The MAP procedures are not easily implemented at this

time. The current means of implementation is by the use of a IBM mainframe, Fortran computer program. This requires that the user not only have access to a mainframe computer, but also must possess some familiarity with executing Fortran programs. Although the current plans to adapt the program for use on personal computers would make these methods more accessible, some elementary knowledge of Fortran would still be required.

III. Performance

The performance of three versions of the minimum average partial correlation method were examined and found to give highly accurate estimations of m. Overall, the largest root of the matrix of partial correlations was the most accurate of the MAP methods, and in some conditions the most accurate of all methods examined. Although this method often had the greatest accuracy, the actual percent of the estimations that were correct was generally the lowest of the three MAP variations. This suggests one might want to examine a range of values for m when using this method. The situations in which the largest root MAP version was less accurate (and more variable) were the high component saturation and the high number of variables per component conditions. In these situations, the method slightly overestimated the number of components.

The trace of the matrix of partial correlations to the

fourth power generally performed better than the original method of the trace of the squared matrix of partial correlations. It was both more accurate and less variable across most conditions, although often the difference between the two methods was very slight. In the optimal conditions of moderate to high component saturation or a high p:m ratio, all three measures of performance for the trace of the matrix of partial correlations to the fourth power were better than the largest root MAP version.

IV. Number of Cases

The three MAP methods were applied to all of the correlation matrices generated in this study. After excluding those cases with a value of m = 0, the three methods estimated the number of components for 85%, 90%, and 96% of the 510 matrices for the trace of the squared matrix of partial correlations, the trace of the partial correlation matrix to the fourth power, and the largest root of the partial correlation matrix, respectively.

The Six Parallel Analysis Methods

I. Rationale

The rationale of parallel analysis is based upon the eigenvalue-greater-than-one rule. Horn (1965) proposed parallel analysis specifically to address the inability of the eigenvalue-greater-than-one rule to reflect sampling error. By

comparing the eigenvalues of the observed data to the eigenvalues of randomly generated data instead of a fixed value of 1.0, random error is taken into account. Although this strengthens the rationale of this method, the criticisms of the eigenvalue-greater-than-one rule rationale still remain applicable to the rationale of parallel analysis.

II. Implementation

Three categories of parallel analysis were examined in this study: regression equations, tabled eigenvalues, and random data generation. The implementation of the methods differs widely. The regression equations may be implemented using available Fortran and Basic programs for IBM compatible personal computers. The estimated eigenvalues for the Allen & Hubbard (1986) equation can be computed using the Basic program presented by Hays (1987) or the Fortran program used in this study (Holden, Longman, Cota, and Fekken, 1989). The latter program also provides estimated eigenvalues for the Longman, Cota, Holden, and Fekken (1989) equation. These programs require some skills or knowledge of computers that may be beyond what many users possess, especially users who are accustomed to using statistical software packages.

The tabled eigenvalues method provided by Lautenschlager (1989a) is the most readily available of all parallel analysis methods. One first locates the table of eigenvalues corresponding to the N and p of the observed data. If there is

no table for the exact value of N or p, the user then selects the values from the N and p tables which are above and below the desired value. Linear interpolation is then conducted on these values to obtain the eigenvalues for the specific values of N or p. Although this method is readily available to all users, it is rather cumbersome to use. If one needs to interpolate across both N and p for a large number of variables, the process can become fairly extensive. At this time, no computer program is known to have been published which would calculate the eigenvalues of this method. The presentation of such a program would greatly simplify the use of this method.

The random data generation method is completed by the use of available computer programs. The Longman, Cota, Holden, and Fekken (1989b) mainframe Fortran program and the IBM compatible personal computer program provided by Lautenschlager (1989b) provide random data eigenvalues. As stated above, these computer programs require more advanced expertise and familiarity with computer systems than is typically observed with many users.

III. Performance

Six variations of parallel analysis were examined. All of the versions produced very accurate estimations of the number of components to retain. In general, the regression equations performed the best of all the parallel analysis methods. All

three equations produced similar results, both in accuracy and in variability, especially when solutions of m = 0 and m = p - 2 were omitted.

Several problems with the regression equations were observed. First, the three equations are limited to the values of N and p from which they were developed. As a result, these methods were evaluated on half of the conditions as most of the other methods.

Second, the Allen & Hubbard (1986) equation and the Lautenschlager, Lance, & Flaherty (1989) equation gave the number of components as p-2 in some of the cases. It is problematic that these two regression equation methods do not give a viable solution in these conditions. A warning to the user that the observed eigenvalues never crossed the random eigenvalues would be helpful for interpreting the solution.

Third, the anomalies of the Allen & Hubbard (1986) equation discussed by Longman, Holden, & Fekken (1991) were observed when p = 48 and N = 300. In this situation, the eigenvalues decreased steadily until the 16th eigenvalue. At the 17th eigenvalue, the eigenvalues began to increase in value and continued to do so until the equation stopped at p - 2. Since the number of components had been indicated at 12 or less components for these cases, this did not interfere with the method giving an accurate solution for m.

Linear interpolation from the tables of mean criterion eigenvalues provided by Lautenschlager (1989a) was also found

to be an accurate method for determining the number of components. Although slightly less accurate than the three regression equations, interpolation of the tabled values does not exhibit the problems of the regression equations. This method was also able to be evaluated across more (390 as compared to 270) of the total 510 combinations of the study conditions than the regression equations.

The random data generation method was the least accurate and most variable of the parallel analysis methods. This finding was surprising since the random data generation method was expected to be the most accurate of the parallel analysis methods. The use of 100 instead of 5 random correlation matrices only marginally improved the accuracy of the method.

IV. Number of Cases

The variations of parallel analysis differed on the number of cases that were examined for each method. The random data generation method was applied with the total number of correlation matrices in this study, 510. After excluding those solutions of m = 0, this method gave estimations for 89% of the matrices.

The tabled eigenvalues method was able to be used on 390 of the 510 correlation matrices. The further reduction for the cases of estimations of m = 0 resulted in this method being employed on 69% of the 510 matrices.

The regression equations were implemented on the fewest

number of the correlation matrices. The three equations were limited to application on 270 of the 510 matrices. After excluding the solutions of m = 0 and m = p - 2, the equations gave viable solutions for 46% to 48% of the correlation matrices. This limitation is a considerable factor in selecting a decision method.

Impact of the Five Factors

The largest effects on all 10 methods were observed with the varied levels of the component saturation and the number of variables per component ratio. These results are consistent with the finding of many other researchers for the effect of component saturation (Guadagnoli & Velicer, 1988; Hakstian, Rogers, & Cattell, 1982; Anderson, Acito, & Lee, 1982; Yeomans & Golder, 1982; Velicer & Fava, 1987; Zwick & Velicer, 1982,1986) and for the effect of the ratio of variables per component (Guadagnoli & Velicer, 1988; Velicer & Fava, 1987; Yeomans & Golder, 1982; Zwick & Velicer, 1982, 1986). Increased accuracy was observed in situations with higher component saturation, with a sharp increase in accuracy observed as the component saturation increased from .40 to .60. The higher variables per component ratio also resulted in increased accuracy.

The third factor manipulated in this study which had a sizable effect on the accuracy of all ten methods was the number of variables. As the number of variables was increased

(24, 48, 72), the variability of all the methods increased and the accuracy decreased.

The methods differed in the effect of the inclusion of unique items. The accuracy of the eigenvalue-greater-than-one rule was decreased by half when unique items were present. The largest root of the matrix of partial correlations showed a very slight decrease in accuracy when the unique items were present. For the rest of the methods, the accuracy increased when unique items were present. These differences were slight for the random data generation methods, the tabled values, and both trace methods.

The ten methods displayed differing patterns of accuracy and variability with the remaining factor varied in this study, sample size. In general, the accuracy of the parallel analysis methods improved as sample size increased. For the minimum average partial correlation methods, as sample size increased the accuracy increased for the trace of the squared matrix of partial correlations, remained about the same for the matrix of partial correlations to the fourth power, and decreased for the largest root of the matrix of partial correlations.

Implications for Future Research

This study examined two promising methods of determining the number of components to retain in a principal components analysis. Both the parallel analysis and the minimum average

partial correlation procedures were found to give accurate estimations of m.

As a first examination of the trace of the matrix of partial correlations to the fourth power and the largest root of the matrix of partial correlations, this study provided evidence of the accuracy of these methods. Further research is needed to substantiate and expand upon these findings. The performance of these methods should be examined with more complex data and more involved conditions. More complex data would include the presence of complex items as well as unique items, and the presence of trivial components (components with less than three variables with substantial loadings). An equal and unequal number of variables per component and of the component saturation of the variables are more complex conditions on which these methods should be evaluated.

Six alternatives of implementing parallel analysis were examined. The more accurate methods were the three regression equations and the tabled values method. These four methods may not be used on data sets with more than 50 variables. Since principal components analysis is a data reduction procedure, commonly used on data sets in excess of 50 variables, this limitation is problematic. Given the accurate performance of these methods, further work to expand the utility of these methods is warranted. The random data generation method was also shown to be an effective method of determining the number of components. Since it is likely to be implemented in those

situations where the regression equations and tabled values are not applicable, further study of the performance of this method is also needed. An examination of the three methods of parallel analysis under the more complex conditions given in the previous paragraph is recommended.

Both the parallel analysis and the MAP methods occasionally retained 0 components. It is recommended that the computer programs employed to implement these methods add warnings to the user in cases of m = 0.

Major Conclusions

1. Of the three methods examined, the MAP methods have the strongest rationale.
2. The eigenvalue-greater-than-one rule is the easiest decision method to implement.
3. The MAP methods, the random data generation parallel analysis method, and the eigenvalue greater-than-one rule were implemented on the total, 510 correlation matrices.
4. The eigenvalue-greater-than-one rule was the least accurate and most variable of all the methods. Continued use of this method is not recommended.
5. The largest root variation of the three MAP methods had the greatest averaged accuracy, lowest variability, but also the lowest percent of correct estimations.
6. The trace of the matrix of partial correlations to the fourth power was the most accurate of the MAP versions in

cases of optimal component identification.

7. The trace of the matrix of partial correlations to the fourth power performed better than the original MAP version of the trace of the squared matrix of partial correlations.
8. Of the parallel analysis methods, the three regression equations displayed the best overall performance.
9. The three regression equations and Lautenschlager's (1989a) tabled eigenvalues method performed better than the random data generation method of parallel analysis.
10. The three factors which had the largest effect on the accuracy of the methods were the component saturation, the variables per component ratio, and the number of variables.
11. All methods performed best with moderate and high component saturation, more variables per component, and fewer variables.

Additional Observations

1. The six parallel analysis methods performed better as sample size increased.
2. The presence of unique items in the data led to improved performance for all methods except the eigenvalue-greater-than-one rule and the largest root MAP version.
3. The six parallel analysis and the three MAP methods

occasionally retained 0 components, usually under conditions of low saturation and the lower p:m ratio.

4. The differences in accuracy, variability, and the percent of correct estimations of m between the random generation parallel analysis method based upon 5 versus 100 replications was trivial.
5. The Allen and Hubbard (1986) and the Lautenschlager, Lance, and Flaherty (1989) regression equations provided estimations of m as p - 2.
6. The Allen and Hubbard (1986) regression equation displayed a previously documented anomaly (Longman, Holden, & Fekken, 1991), but it did not affect the accuracy of the estimation of m.
7. The three regression equations and Lautenschlager's (1989a) tabled eigenvalues method are limited to data sets with 5 to 50 variables.
8. The Allen and Hubbard (1986) equation may be used with data sets with a maximum of 1000 subjects, the Longman, Cota, Holden, and Fekken (1989) equation with 500 subjects, and the Lautenschlager, Lance, and Flaherty (1989) equation with 1000 subjects. Lautenschlager's (1989a) tabled eigenvalues method is applicable for data sets with up to 2000 subjects.
9. It is recommended that warnings be added to the computer programs used to implement these methods for solutions of m = 0 or m = p - 2.

Table 1
Overall Design of the Study

CS	P	P:M	M	N
Part I. Ideal Patterns Only				
	24	8:1	3	75 150 300
		4:1	6	
.40	48	8:1	6	
		4:1	12	
	72	8:1	9	
		4:1	18	
	24	8:1	3	
		4:1	6	
.60	48	8:1	6	
		4:1	12	
	72	8:1	9	
		4:1	18	
	24	8:1	3	
		4:1	6	
.80	48	8:1	6	
		4:1	12	
	72	8:1	9	
		4:1	18	
Part II. Ideal Patterns and Unique Items				
	27 (3 unique)	8:1	3	75 150 300
	30 (6 unique)	4:1	6	
.40	54 (6 unique)	8:1	6	
	60 (12 unique)	4:1	12	
	81 (9 unique)	8:1	9	*
	90 (18 unique)	4:1	18	*
	27 (3 unique)	8:1	3	
	30 (6 unique)	4:1	6	
.60	54 (6 unique)	8:1	6	
	60 (12 unique)	4:1	12	
	81 (9 unique)	8:1	9	*
	90 (18 unique)	4:1	18	*
	27 (3 unique)	8:1	3	
	30 (6 unique)	4:1	6	
.80	54 (6 unique)	8:1	6	
	60 (12 unique)	4:1	12	
	81 (9 unique)	8:1	9	*
	90 (18 unique)	4:1	18	*

* No cases were generated when N < P

Table 2
Sequence of Population Correlation Matrix Generation

I. Component Pattern Matrix

$$A = \begin{matrix} .60 & 00 \\ .60 & 00 \\ .60 & 00 \\ 00 & .60 \\ 00 & .60 \\ 00 & .60 \end{matrix}$$

II.

$$R_1 = A\,A' = \begin{matrix} .36 & .36 & .36 & 00 & 00 & 00 \\ .36 & .36 & .36 & 00 & 00 & 00 \\ .36 & .36 & .36 & 00 & 00 & 00 \\ 00 & 00 & 00 & .36 & .36 & .36 \\ 00 & 00 & 00 & .36 & .36 & .36 \\ 00 & 00 & 00 & .36 & .36 & .36 \end{matrix}$$

III.

$$D^2 = \begin{matrix} .64 & 00 & 00 & 00 & 00 & 00 \\ 00 & .64 & 00 & 00 & 00 & 00 \\ 00 & 00 & .64 & 00 & 00 & 00 \\ 00 & 00 & 00 & .64 & 00 & 00 \\ 00 & 00 & 00 & 00 & .64 & 00 \\ 00 & 00 & 00 & 00 & 00 & .64 \end{matrix}$$

IV. Population Correlation Matrix

$$R = R_1 + D^2 = \begin{matrix} 1.00 & .36 & .36 & 00 & 00 & 00 \\ .36 & 1.00 & .36 & 00 & 00 & 00 \\ .36 & .36 & 1.00 & 00 & 00 & 00 \\ 00 & 00 & 00 & 1.00 & .36 & .36 \\ 00 & 00 & 00 & .36 & 1.00 & .36 \\ 00 & 00 & 00 & .36 & .36 & 1.00 \end{matrix}$$

Table 3
Deviation Score Means, Standard Deviations, Percent of Correct Estimations, and Number of Estimations Collapsed Across All Conditions

	Mean	S.D.	Percent Accurate	N
I. Parallel Analysis Procedures				
AH	-1.16	8.87	80.7	270
LCHF	1.03	2.64	80.0	270
LLF	.60	3.87	80.0	270
R5	2.13	4.59	72.0	510
R100	2.11	4.63	72.7	510
TAB	1.50	3.86	79.5	390
II. Minimum Average Partial Correlation Procedures				
TR2	2.46	4.83	66.9	510
TR4	1.89	4.48	72.5	510
LR	.59	4.03	45.9	510
III. Eigenvalues-Greater-Than-One Rule				
	-6.27	5.85	22.0	510

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 4
Deviation Score Means, Standard Deviations, and Percent of Correct Estimations By Three Levels of Component Saturation

	Component Saturation								
	.40			.60			.80		
	Mean	S.D.	%	Mean	S.D.	%	Mean	S.D.	%
I. Parallel Analysis Procedures									
AH	-1.54	13.21	51.1	-1.94	7.82	91.1	.00	.00	100.0
LCHF	2.83	3.92	52.2	.21	.81	91.1	.03	.18	96.7
LLF	1.39	6.51	54.4	.33	1.36	91.1	.07	.29	94.4
R5	5.58	6.47	40.6	.68	1.79	81.2	.14	.63	94.1
R100	5.56	6.57	42.4	.65	1.70	81.2	.14	.64	94.7
TAB	4.25	5.74	51.5	.25	.80	88.5	.02	.12	98.5
II. Minimum Average Partial Correlation Procedures									
TR2	6.56	6.18	20.6	.85	2.50	82.4	-.01	.15	97.6
TR4	5.66	6.20	38.2	.09	.74	87.6	-.09	.30	91.8
LR	3.49	5.65	24.1	-.92	1.42	50.0	-.80	1.50	63.5
III. Eigenvalues-Greater-Than-One Rule									
	-11.22	5.83	00.0	-5.36	4.43	15.3	-2.22	2.75	50.6

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 5
Deviation Score Means, Standard Deviations, and Percent of Correct Estimations By Two Levels of the Variables:Component Ratio

	Variables Per Component Ratio 4:1			8:1		
	Mean	S.D.	%	Mean	S.D.	%
I. Parallel Analysis Procedures						
AH	-.85	10.01	65.9	-1.47	7.58	95.6
LCHF	1.95	3.47	65.9	.10	.46	94.1
LLF	1.01	5.39	65.9	.18	.85	94.1
R5	3.86	5.87	55.7	.40	1.35	88.2
R100	3.85	5.92	56.1	.38	1.32	89.4
TAB	2.87	5.07	64.1	.14	.73	94.9
II. Minimum Average Partial Correlation Procedures						
TR2	4.40	6.10	53.7	.52	1.39	80.0
TR4	3.70	5.75	58.4	.08	.80	86.7
LR	2.06	5.06	36.1	-.88	1.62	55.7
III. Eigenvalues-Greater-Than-One Rule						
	-5.83	5.19	23.5	-6.71	6.42	20.4

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 6
Deviation Score Means, Standard Deviations, and Percent of Correct Estimations By Three Levels of the Number of Variables

	Number of Variables								
	24			48			72		
	Mean	S.D.	%	Mean	S.D.	%	Mean	S.D.	%
I. Parallel Analysis Procedures									
AH	.62	1.76	85.0	-4.72	14.57	72.2	*	*	*
LCHF	.66	1.81	86.1	1.76	3.69	67.8	*	*	*
LLF	-.12	3.66	85.6	2.03	3.91	68.9	*	*	*
R5	.59	1.69	86.7	2.21	4.10	66.1	3.89	6.52	61.3
R100	.56	1.62	86.7	2.19	4.14	68.3	3.89	6.57	61.3
TAB	.56	1.64	87.2	1.97	4.02	72.7	3.18	6.58	73.3
II. Minimum Average Partial Correlation Procedures									
TR2	1.14	2.23	73.3	2.54	4.49	64.4	3.95	6.71	62.0
TR4	.94	2.13	82.2	1.97	4.34	70.6	2.93	6.20	63.3
LR	.48	1.76	72.8	.63	4.14	36.1	.67	5.59	25.3
III. Eigenvalues-Greater-Than-One Rule									
	-2.86	2.39	27.2	-6.76	4.96	20.6	-9.79	7.33	17.3

* No estimations can be generated for this condition

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 7
Deviation Score Means, Standard Deviations, and Percent of Correct Estimations By Three Levels of Sample Size

	Sample Size								
	75			150			300		
	Mean	S.D.	%	Mean	S.D.	%	Mean	S.D.	%
I. Parallel Analysis Procedures									
AH	-5.47	13.88	67.8	1.27	3.14	83.3	.71	2.63	91.1
LCHF	1.67	3.08	64.4	1.30	3.15	81.1	.11	.53	94.4
LLF	1.93	3.29	61.1	.06	5.11	83.3	-.20	2.37	95.6
R5	2.96	4.47	52.0	2.23	4.86	71.7	1.34	4.30	88.9
R100	2.91	4.44	53.3	2.18	4.86	72.2	1.38	4.45	89.4
TAB	1.72	3.16	65.6	1.80	4.36	76.7	1.07	3.71	90.7
II. Minimum Average Partial Correlation Procedures									
TR2	2.99	4.66	52.0	2.46	4.87	66.1	2.04	4.90	80.0
TR4	1.89	4.12	56.0	1.81	4.45	75.6	1.96	4.82	83.3
LR	.39	3.67	34.0	.54	4.44	42.2	.81	3.89	59.4
III. Eigenvalues-Greater-Than-One Rule									
	-6.12	5.06	18.0	-6.88	6.25	21.1	-5.78	6.04	26.1

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 8
Deviation Score Means, Standard Deviations, and Percent of Correct Estimations By the Presence of Unique Items

	Present			Absent		
	Mean	S.D.	%	Mean	S.D.	%
I. Parallel Analysis Procedures						
AH	.64	1.84	82.2	-2.07	10.68	80.0
LCHF	.69	1.83	84.4	1.19	2.95	77.8
LLF	-.24	4.37	83.3	1.02	3.54	78.3
R5	2.06	4.52	73.3	2.19	4.66	70.7
R100	2.08	4.55	72.9	2.14	4.70	72.6
TAB	1.17	3.04	82.0	1.71	4.29	77.9
II. Minimum Average Partial Correlation Procedures						
TR2	2.25	4.61	66.7	2.66	5.01	67.0
TR4	1.76	4.34	72.9	2.00	4.62	72.2
LR	.62	3.62	52.5	.56	4.37	40.0
III. Eigenvalues-Greater-Than-One Rule						
	-8.49	5.59	0.0	-4.30	5.36	41.5

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 9
Averaged Values of M Retained by the Eigenvalue-Greater-Than-One Rule, Overall and For Each Level of Component Saturation and the P:M Ratio

	Overall	Component Saturation			P:M Ratio	
		.40	.60	.80	4:1	8:1
24	7.4	9.8	6.5	5.8	8.9	5.8
48	15.8	20.9	14.8	11.6	18.3	13.2
72	23.3	31.0	22.4	16.5	26.7	19.8

Table 10
Number and Percent of Estimations Which Overestimated and Underestimated the Value of M, Collapsed Across All Factors

	Over		Under	
	N	%	N	%
I. Parallel Analysis				
AH	18	6.7	34	12.6
LCHF	0	0.0	54	20.0
LLF	8	2.9	46	17.0
R5	0	0.0	143	28.0
R100	0	0.0	139	27.2
TAB	0	0.0	80	20.5
II. Minimum Average Partial Correlation Procedures				
TR2	4	.8	165	32.4
TR4	26	5.1	114	22.4
LR	184	36.1	92	18.0
III. Eigenvalues-Greater-Than-One Rule				
	398	78.0	0	0.0

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 11
Percent of Estimations Which Overestimated and Underestimated the Value of M, By Three Levels of Component Saturation

	Component Saturation					
	.40		.60		.80	
	Over	Under	Over	Under	Over	Under
I. Parallel Analysis Procedures						
AH	11.2	37.7	8.9	0.0	0.0	0.0
LCHF	0.0	47.8	0.0	8.8	0.0	3.3
LLF	7.7	37.7	1.1	7.7	0.0	5.5
R5	0.0	59.4	0.0	18.8	0.0	5.9
R100	0.0	57.6	0.0	18.8	0.0	5.3
TAB	0.0	48.5	0.0	11.5	0.0	1.5
II. Minimum Average Partial Correlation Procedures						
TR2	0.0	79.4	.6	17.0	1.8	.6
TR4	1.8	60.0	5.3	7.0	8.2	0.0
LR	24.7	51.2	47.0	3.0	36.5	0.0
III. Eigenvalues-Greater-Than-One Rule						
	100.0	0.0	84.7	0.0	49.4	0.0

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 12
Percent of Estimations Which Overestimated and Underestimated the Value of M, By Two Levels of the Variables:Component Ratio

	Number of Variables Per Component 4:1		8:1	
	Over	Under	Over	Under
I. Parallel Analysis Procedures				
AH	9.6	24.4	3.7	.7
LCHF	0.0	34.1	0.0	6.0
LLF	5.9	28.1	0.0	5.9
R5	0.0	44.3	0.0	11.8
R100	0.0	43.9	0.0	10.6
TAB	0.0	35.9	0.0	5.1
II. Minimum Average Partial Correlation Procedures				
TR2	1.2	45.1	.4	19.6
TR4	3.5	38.0	6.7	6.7
LR	30.6	33.3	41.5	2.8
III. Eigenvalues-Greater-Than-One Rule				
	76.5	0.0	79.6	0.0

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 13
Percent of Estimations Which Overestimated and Underestimated the Value of M, By Three Levels of the Number of Variables

	Number of Variables					
	24		48		72	
	Over	Under	Over	Under	Over	Under
I. Parallel Analysis Procedures						
AH	1.7	13.3	16.7	11.1	*	*
LCHF	0.0	13.9	0.0	32.2	*	*
LLF	4.5	10.1	0.0	31.1	*	*
R5	0.0	13.3	0.0	33.9	0.0	38.7
R100	0.0	13.3	0.0	31.7	0.0	38.7
TAB	0.0	12.8	0.0	27.3	0.0	26.7
II. Minimum Average Partial Correlation Procedures						
TR2	0.0	26.7	2.2	33.3	0.0	38.0
TR2	.6	17.3	6.1	23.3	9.3	27.3
LR	13.4	13.9	45.1	18.9	52.7	22.0
III. Eigenvalues-Greater-Than-One Rule						
	72.8	0.0	79.4	0.0	82.7	0.0

* No estimations can be generated for this condition

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 14
The Number of Occurrences Where the Regression Equation Methods Give the Number of Components as P - 2

Equation	Number of Occurrences	P	N	CS	P:M	U
AH	5	48	75	.4	8:1	Absent
AH	5	48	75	.4	4:1	Absent
AH	5	48	75	.6	4:1	Absent
LLF	2	24	150	.4	4:1	Absent
LLF	2	24	150	.4	4:1	Present
LLF	1	24	300	.4	4:1	Present

AH = Allen and Hubbard (1986) regression equation
LLF = Lautenschlager et al (1989) regression equation

Table 15
Deviation Score Means, Standard Deviations, Percent of Correct Estimations, and Number of Estimations for the Three Regression Equations When Estimations of M = P - 2 Are Omitted, Collapsed Across All Factors

	Mean	S.D.	Percent Accurate	N
AH	.89	2.64	85.5	255
LCHF	1.03	2.64	80.0	270
LLF	.98	2.69	81.5	265

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation

Table 16
Deviation Score Means, Standard Deviations, and Percent of Correct Estimations for the Three Regression Equations When Estimations of M = P - 2 Are Omitted, By Component Saturation, P:M Ratio, and the Number of Variables

		Component Saturation			P:M Ratio		Number of Variables		
		.40	.60	.80	4:1	8:1	24	48	72
AH	M	2.89	-.06	.00	1.80	.01	.62	1.53	*
	SD	4.05	.32	.00	3.55	.09	1.76	3.97	*
	%	57.50	96.50	100.00	71.20	99.20	85.00	86.70	*
LCHF	M	2.83	.21	.03	1.95	.10	.66	1.76	*
	SD	3.92	.81	.18	3.47	.46	1.81	3.69	*
	%	52.20	91.10	96.70	65.90	94.10	86.10	67.80	*
LLF	M	2.62	.33	.07	1.80	.18	.43	2.03	*
	SD	4.08	1.36	.29	3.57	.85	1.51	3.91	*
	%	57.60	91.10	94.40	68.50	94.10	88.00	68.90	*

* No cases were generated for this condition

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation

Table 17
Number and Percent of Occurrences of M = 0, Collapsed Across All Factors

	N	%
I. Parallel Analysis Procedures		
AH	20	7.8
LCHF	24	8.9
LLF	22	8.3
R5	57	11.2
R100	57	11.2
TAB	37	9.5
II. Minimum Average Partial Correlation Procedures		
TR2	77	15.1
TR4	50	9.8
LR	22	4.3

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Table 18
Deviation Score Means, Standard Deviations, Percent of Correct Estimations, and Number of Estimations When Estimations of M = P - 2 or M = 0 Are Omitted, Collapsed Across All Factors

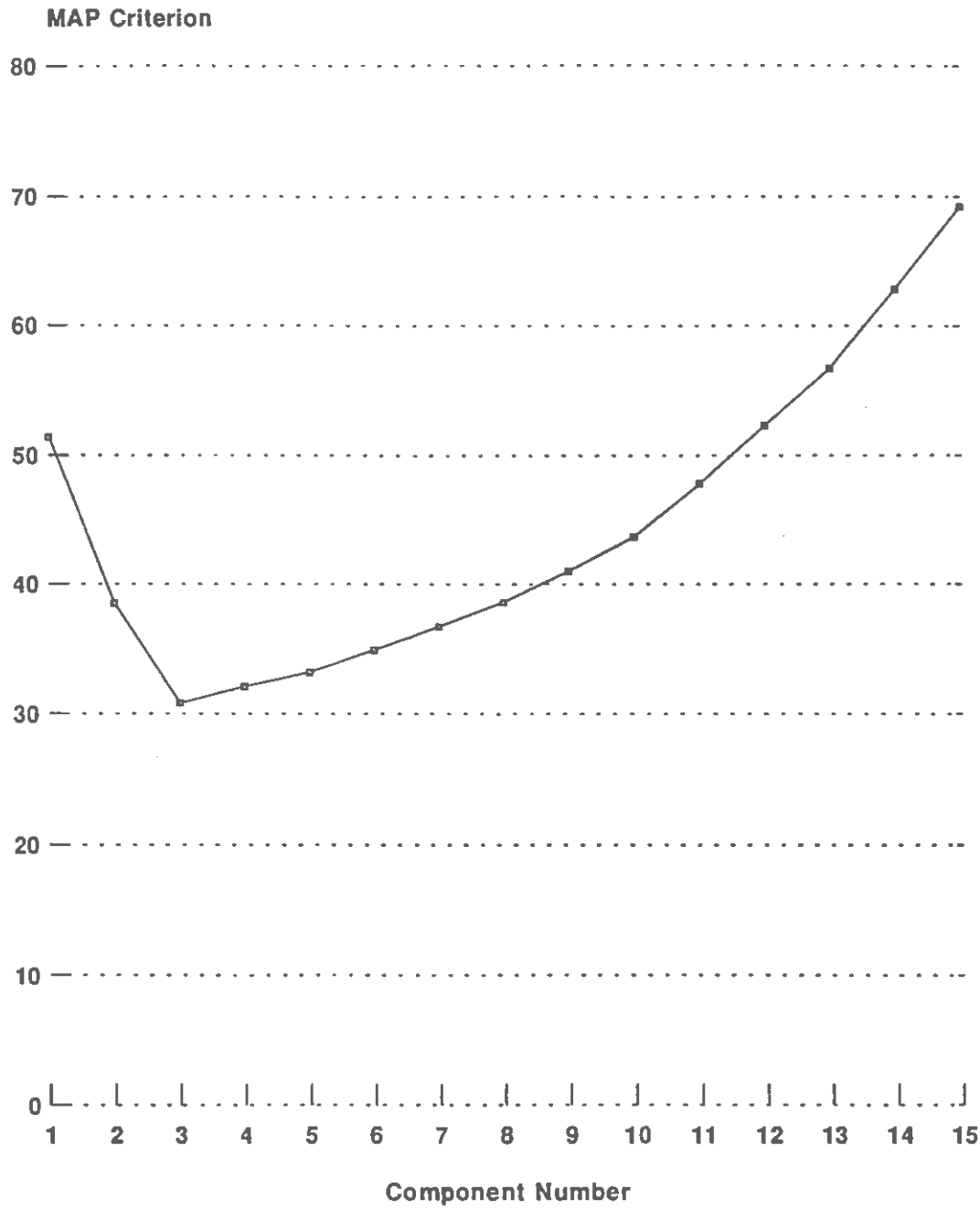
	Mean	S.D.	Percent Accurate	N
I. Parallel Analysis Procedures				
AH	.22	1.07	92.8	235
LCHF	.35	1.27	87.8	246
LLF	.35	1.23	87.8	246
R5	.77	2.16	81.2	452
R100	.75	2.18	81.9	453
TAB	.49	1.88	87.8	353
II. Minimum Average Partial Correlation Procedures				
TR2	.93	2.75	78.8	433
TR4	.99	3.52	80.6	459
LR	.21	3.64	48.0	488

* No cases were generated for this condition

AH = Allen and Hubbard (1986) regression equation
LCHF = Longman et al (1989) regression equation
LLF = Lautenschlager et al (1989) regression equation
R5 = Generation of 5 random data correlation matrices
R100 = Generation of 100 random data correlation matrices
TAB = Lautenschlager (1989) tabled eigenvalues
TR2 = Trace, partial correlation matrix, second power
TR4 = Trace, partial correlation matrix, fourth power
LR = Largest root, partial correlation matrix

Figure 1

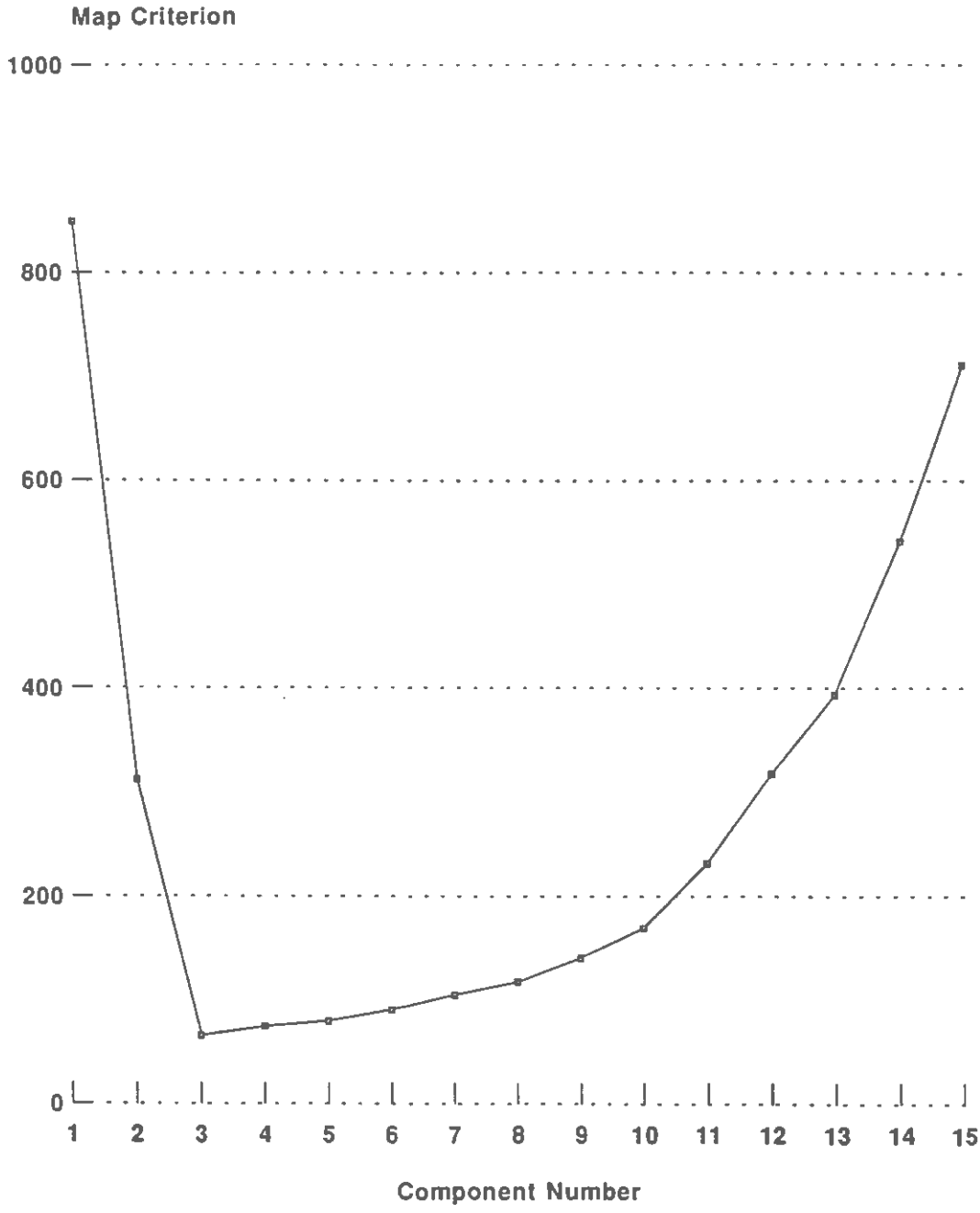
Minimum Average Partial Correlation



Trace of the Squared Matrix
M = 3, P = 24, N = 75, CS = .60

Figure 2

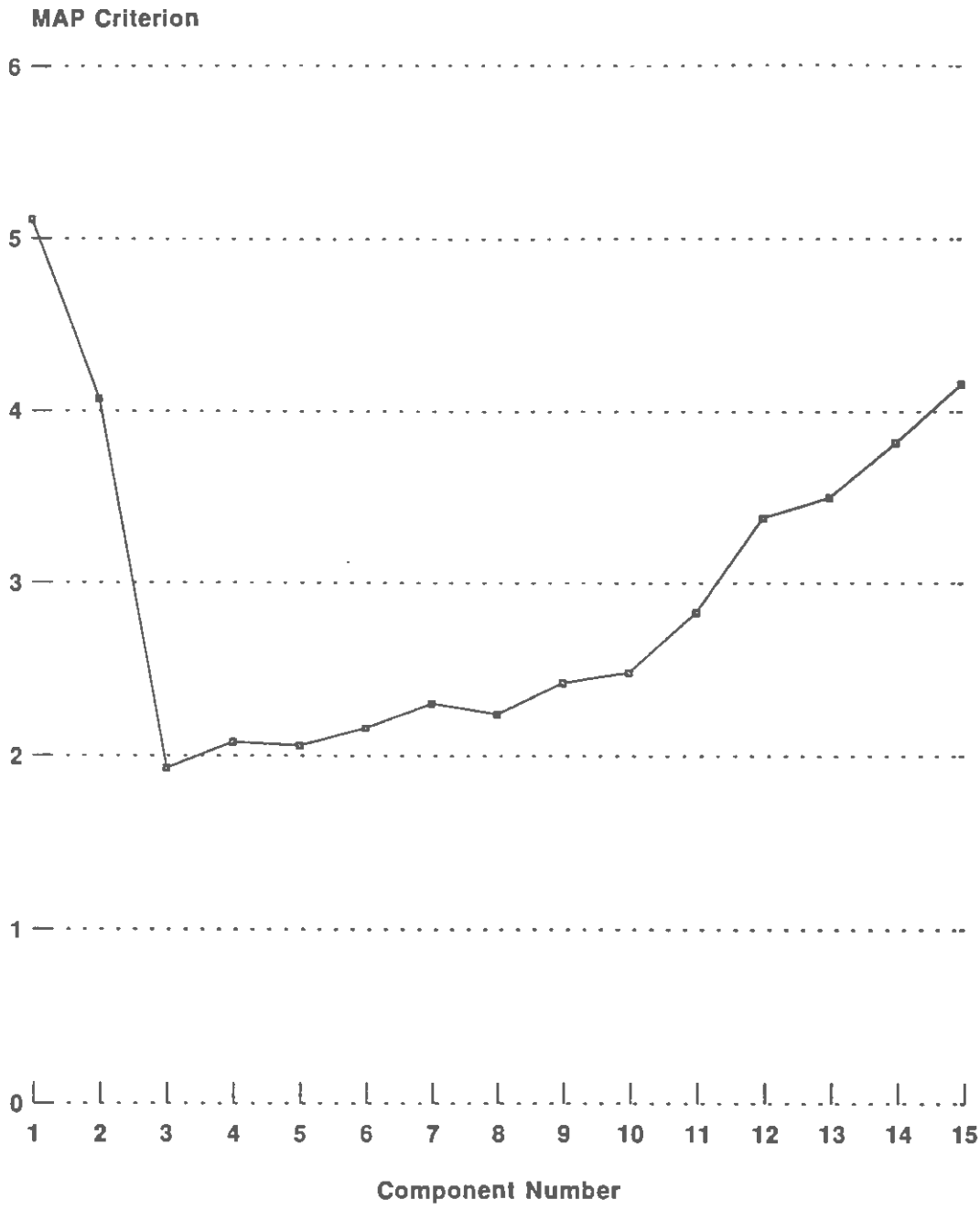
Minimum Average Partial Correlation



Trace of the Matrix to the Fourth Power
M = 3, P = 24, N = 75, CS = .60

Figure 3

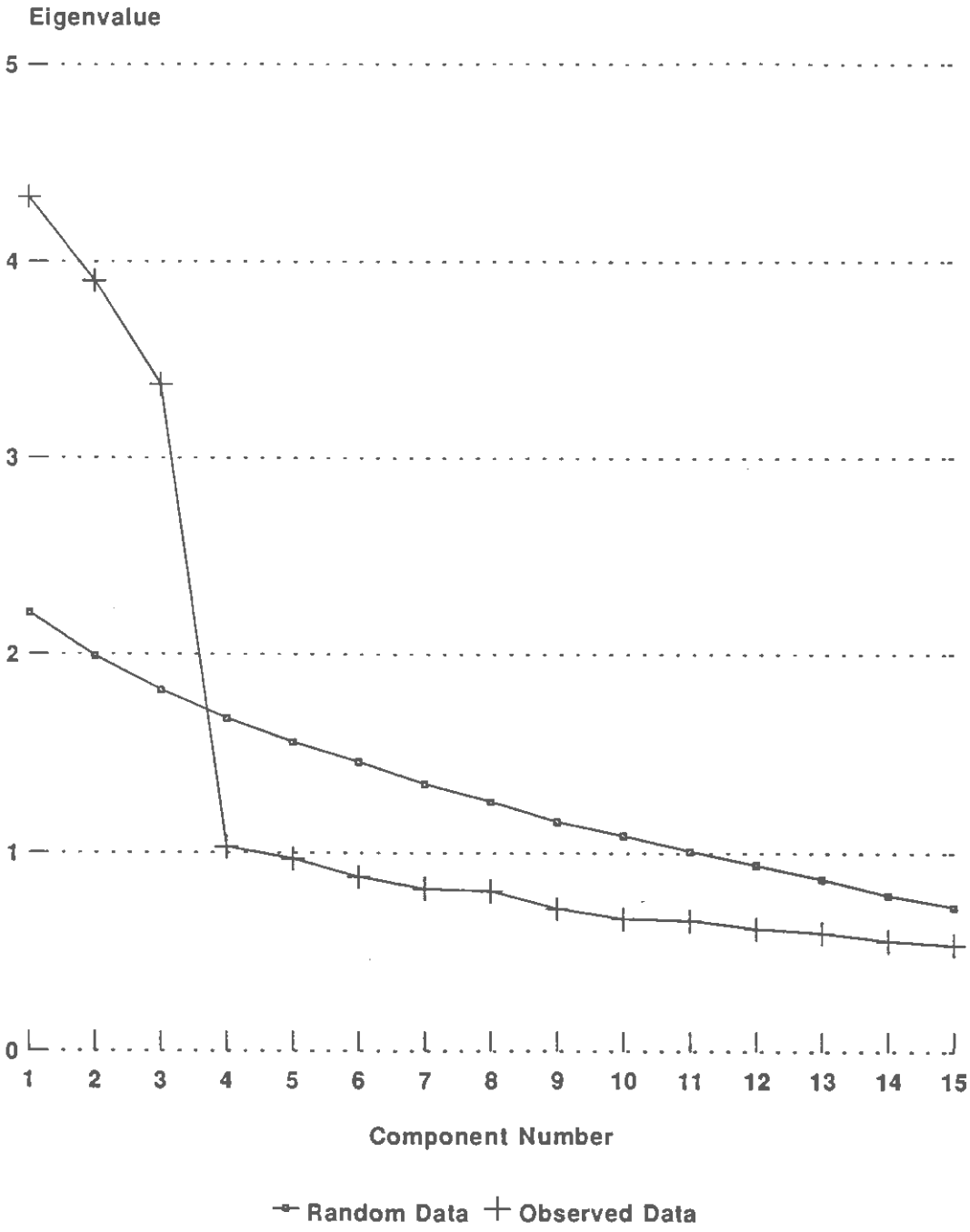
Minimumn Average Partial Correlation



Largest Root of the Matrix
M = 3, P = 24, N = 75, CS = .60

Figure 4

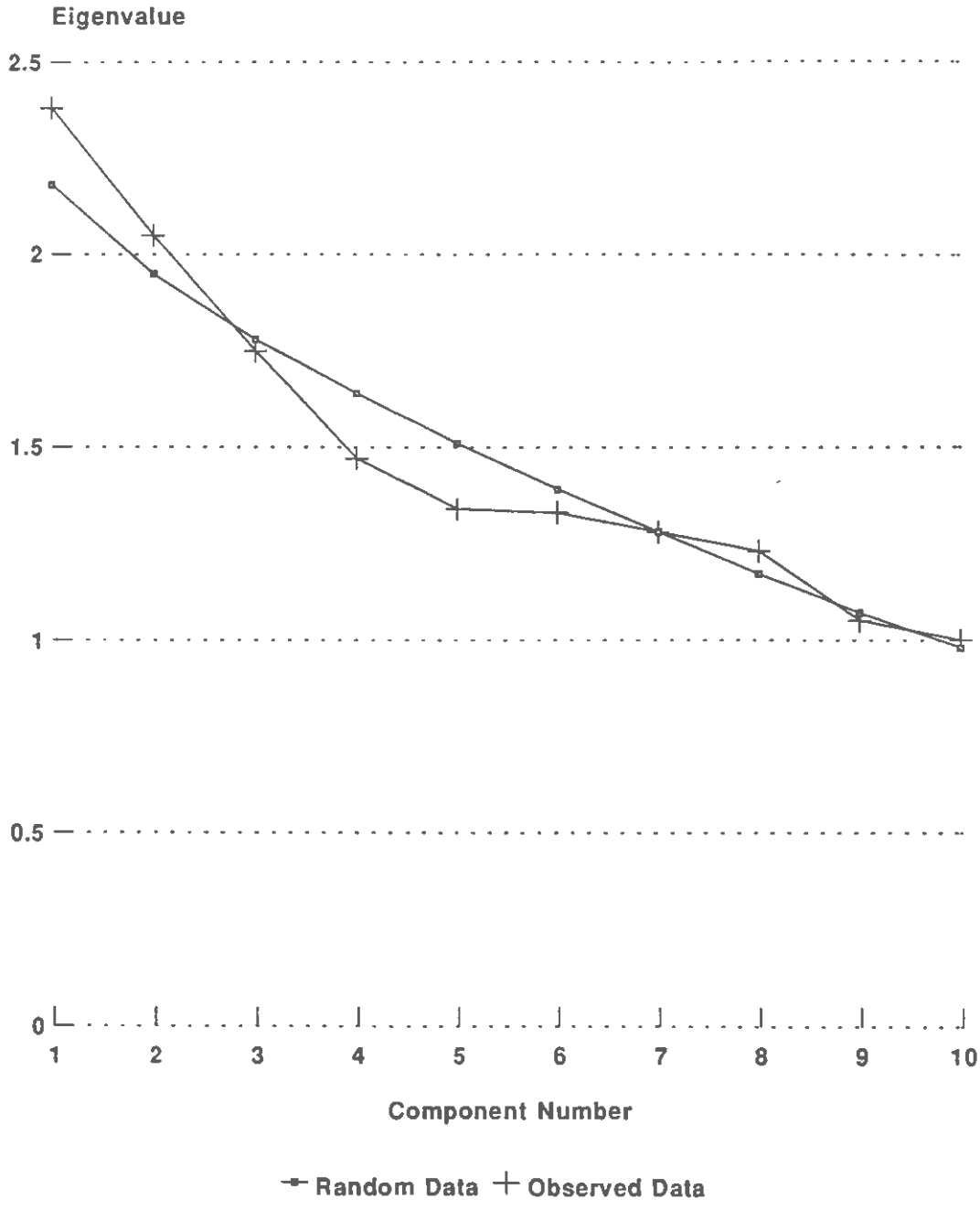
Parallel Analysis



Random Generation, 100 Matrices
M = 3, P = 24, N = 75, CS = .60

Figure 5

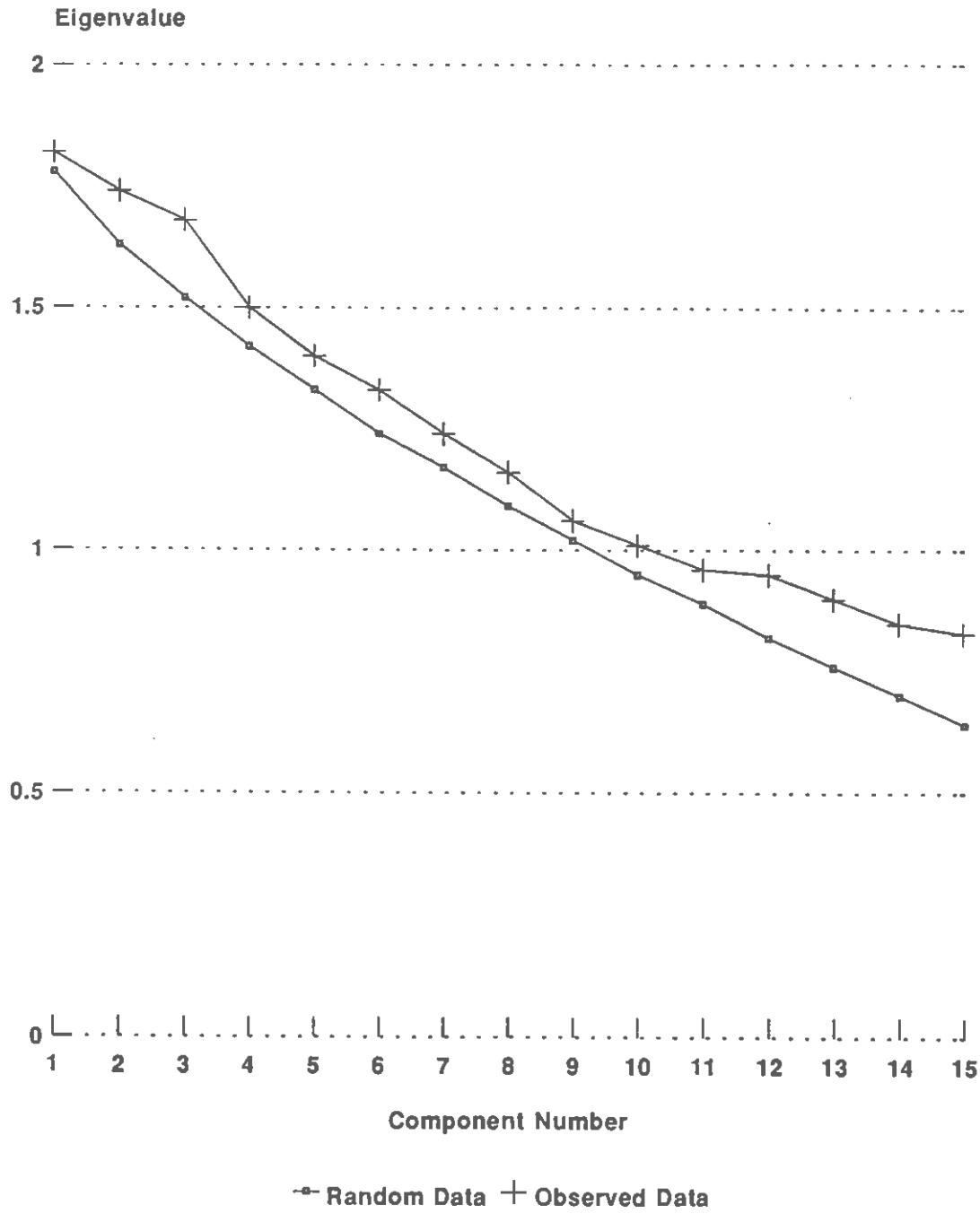
Parallel Analysis



Alternating Higher Value
Allen & Hubbard Equation
M = 3, P = 24, N = 75, CS = .40

Figure 6

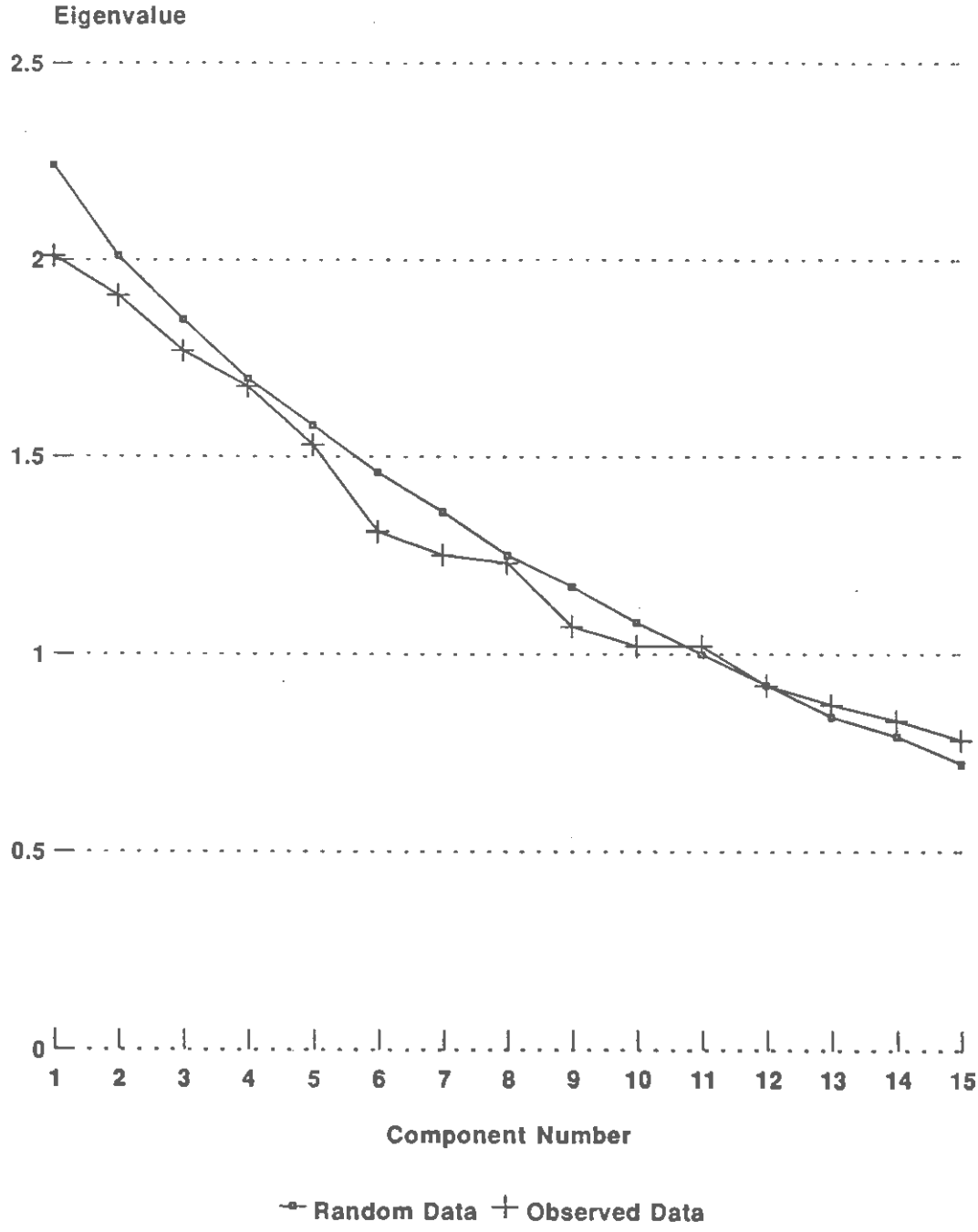
Parallel Analysis: M = P - 2



Lautenschlager et al. Equation
M = 6, P = 24, N = 150, CS = .40

Figure 7

Parallel Analysis: M = 0



Longman et al. Equation
M = 6, P = 24, N = 75, CS= .40

BIBLIOGRAPHY

Allen, S. J., & Hubbard, R. (1986). Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. Multivariate Behavioral Research, 21, 393-398.

Anderson, R. D., Acito, F. & Lee, H. (1982). A simulation study of three methods for determining the number of image components. Multivariate Behavioral Research, 17, 493-502.

Bartlett, M. S. (1950). Tests of significance in factor analysis. British Journal of Psychology, Statistical Section, 3, 77-85.

Bartlett, M. S. (1951). A further note on tests of significance in factor analysis. British Journal of Psychology, Statistical Section, 4, 1-2.

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

Cattell, R. B. & Jaspers, J. (1967). A general plasmode for factor analytic exercises and research. Multivariate Behavioral Research Monographs, 3, 1-212.

Cattell, R. B. & Vogelmann, S. A. (1977). A comprehensive trial of the scree and Kaiser-Guttman criteria for determining the number of factors. Multivariate Behavioral Research, 12, 289- 325.

Cliff, N. (1970). The relation between sample and population characteristic vectors. Psychometrika, 35, 163-178.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. Psychological Bulletin, 103, 276.279.

Cliff, N. & Hamburger, C. (1967). Study of sampling errors in factor analysis by means of artificial experiments. Psychological Bulletin, 68, 430-445.

Crawford, C. B. (1975). Determining the number of interpretable factors. Psychological Bulletin, 82, 226-237.

Crawford, C. B., & Koopman, P. (1973). A note on Horn's test for the number of factors in factor analysis. Multivariate Behavioral Research, 8, 117-125.

Everett, J. E. (1983). Factor comparability as a means of determining the number of factors and their rotation. Multivariate Behavioral Research, 18, 197-218.

Gorsuch, R.L. (1973). Using Bartlett's significance test to determine the number of factors to extract. Educational and Psychological Measurement, 33, 361-364.

Gorsuch, R. L. (1983). Factor Analysis. (2nd ed.) Hillsdale, N.J.: Lawrence Erlbaum.

Guttman, L. (1954). Some necessary conditions for common factor analysis. Psychometrika, 19, 149-162.

Guadagnoli, E. & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. Psychological Bulletin, 103, 265-275.

Guadagnoli, E. & Velicer, W. (1991). A comparison of pattern

matching indices. Multivariate Behavioral Research, 26, 323-343.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. Multivariate Behavioral Research, 17, 193-219.

Hays, R. D. (1987). PARALLEL: A program for performing parallel analysis. Applied Psychological Measurement, 11, 58.

Holden, R. R., Longman, R. S., Cota, A. A., & Fekken, G. C. (1989). PAR: Parallel analysis routine for random data eigenvalue estimation. Applied Psychological Measurement, 13, 192.

Horn, J. L. (1965). A Rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Horn, J. L. & Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. Multivariate Behavioral Research, 14, 283-300.

Hubbard, R. & Allen, S. J. (1987). An empirical comparison of alternate methods for principal components extraction. Journal of Business Research, 15, 173-190.

Humphreys, L. G. & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behavioral Research, 10, 193-205.

Kaiser, H. F. (1960). The application of electronic computers

to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Lautenschlager, G. J. (1989a). A comparison of alternatives to conducting monte carlo analyses for determining parallel analysis criteria. Multivariate Behavioral Research, 24, 365- 395.

Lautenschlager, G. J. (1989b). PARANAL.TOK: A program for developing parallel analysis criteria. Applied Psychological Measurement, 13, 176.

Lautenschlager, G. J., Lance, C. E., & Flaherty, V. L. (1989). Parallel analysis criteria: Revised equations for estimating the latent roots of random data correlation matrices. Educational and Psychological Measurement, 49, 339-345.

Lee, H. B. & Comrey, A. L. (1979). Distortions in a commonly used factor analytic procedure. Multivariate Behavioral Research, 14, 301-321.

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. Psychometrika, 33, 37-71.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989a). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. Multivariate Behavioral Research, 24, 59-69.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989b). PAM: A double precision FORTRAN routine for the

parallel analysis method in principle components analysis. Behavior Research Methods, Instruments, and Computers, 21, 477-480.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1991). Implementing parallel analysis for principal components analysis: A comparison of six methods. Manuscript under review.

Longman, R. S., Holden, R. R., & Fekken, G. C. (1991). Anomalies in the Allen and Hubbard parallel analysis procedure. Applied Psychological Measurement, 15, 95-97.

Montanelli, Jr., R. G. (1975). A computer program to generate sample correlation and covariance matrices. Educational and Psychological Measurement, 35, 195-197.

Montanelli, Jr., R. G. & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. Psychometrika, 41, 341-348.

Odell, P. L. & Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. Journal of the American Statistical Association, 61, 199-203.

Revelle, W. & Rocklin, T. (1979). Very simple structure: an alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behavioral Research, 14, 403-414.

Schonemann, P. H. (1990). Facts, fictions, and common sense about factors and components. Multivariate Behavioral

Research, 25, 47-51.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. Psychometrika, 34, 421-459.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika, 31, 321-327.

Velicer, W. F. & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. Multivariate Behavioral Research, 22, 193-209.

Velicer, W. F., Fava, J. L., Zwick, W. R., & Harrop. J. W. (1990). CAX (Computer program). University of Rhode Island, Kingston.

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. Multivariate Behavioral Research, 17, 371-388.

Yeomans, K. A. & Golder, P. A. (1982). The Guttman-Kaiser criterion as a predictor of the number of common factors. The Statistician, 31, 221-229.

Zwick, W. R., & Velicer W. F. (1982). Factors influencing four rules for determining the number of components to retain. Multivariate Behavioral Research, 17, 253-269.

Zwick, W. R., & Velicer W. F. (1986). Comparison of five rules

for determining the number of components to retain. Psychological Bulletin, 99, 432-442.